REMARKS/ARGUMENTS

Claims 20 - 37 have been amended as a consequence of the *vacatur* and of the interview courteously granted to applicant and his undersigned attorney by the Examiner on November 29, 2006.

At the interview, after initially discussing the effect of the *vacatur* on the prosecution, applicant and the Examiner engaged in a claim construction analysis addressing issues raised in the *vacatur* and the PTO "Guidelines" that resulted in the claim amendments submitted herewith.

With respect to the main issue of biological experimentation and the provision of the physical experimental proof heretofore required by the Examiner, and the computational methods utilized by the applicant herein, the Board, in its *vacatur*, noted that the rejection based on 35 U.S.C. 112, first paragraph, "suffered from several deficiencies." First, the *vacatur* referred to claim 21 and pointed out that claim 21 requires "detecting, by computer, changes in connectron behavior in the genome as a function of changes in the sequence of the genome." And the *vacatur* further pointed out that:

> [T]he rejection, however, does not address the limitations of all of the independent claims, nor does it address the limitations of dependent claims 37.
>
> Second, the rejection appears to address limitations that do not seem to appear in the claims. The examiner focuses on the issue that "[i]n order to practice the claimed invention one of skill in the art must identify and use a connectron to predict regulation of gene expression."

The *vacatur* noted that claim 20 did not require predicting regulation of gene expression, but only appears to require locating possible connectrons.

Applicant respectfully submits that this is in effect a repudiation of the Examiner's 35 U.S.C. 112, first paragraph, requirement that applicant submit experimental biological evidence and that the *vacatur* sustained the applicant's position that

applicant's computational methods satisfy 35 U.S.C. 112 enablement requirements. (See the attached paper by the applicant entitled "Further support of the Applicant's position that computational methods do not require physical proof to support patentability" and the papers attached thereto.) All of the claims are directed to computer mediated methods and do not require biological experimentation to sustain them.

In view of the extensive amendments to the claims to bring them into conformance with the *vacatur* and guidelines set out by the Patent Office, it is respectfully submitted that claims 20 - 37 are in condition for allowance; and further and favorable action is requested.

Respectfully submitted,

Jim Zegeer

Jim Zegeer, Reg. No. 18,957
Attorney for Applicant

Attachments:
"Further support of the Applicant's position that computational methods do not require physical proof to support patentability" and Addendums #1 - #7.

Suite 108
801 North Pitt Street
Alexandria, VA 22314
Telephone: 703-684-8333

Date: December 6, 2006

In the event this paper is deemed not timely filed, the applicant hereby petitions for an appropriate extension of time. The fee for this extension may be charged to Deposit Account No. 26-0090 along with any other additional fees which may be required with respect to this paper.

Further support of the Applicant's position that computational methods do not require physical proof to support patentability

This section is a collection of seven addenda and supporting commentary. Each addendum is a document available in the public media.

Addendum 1

J. David Rawn is professor of biochemistry and bioinformatics at Towson University in Towson, Maryland. Dr. Rawn is the author of a number of biochemistry textbooks and has a bioinformatics book in preparation.

In this essay Dr. Rawn establishes that computation should be used to form theories so as to enable efficient physical experimentation. He sees a cycle of hypothesis formation, followed by initial experimentation, and then a phase of model refinement.

Addendum 2

Dr. Gary Peltz working at Roche Laboratories in Palo Alto, California describes a New Genomic Method [that] "can identify disease-causing genes with unprecedented precision and speed" - http://www.roche.com/med-cor-2004-10-22b . Peltz says that "our hope is that this new computational approach will increase the utility of the vast amount of DNA sequence information available today and help researchers more fully leverage mouse models of human disease to identify genes contributing to disease risk and drug response". The original methods paper was published in Science in June of 2001. The method of Peltz et al has now been applied to the mouse genome to produce predicted regions containing hundreds of genes and the results are assessed by relative statistical criteria.

The importance of this Roche work is that the time difference between the development of the methodology and its application to biomedical problems takes five years that roughly correspond to the development of the Connectron methodology and its more recent application to the mouse transcriptome. The Roche methodology identifies regions of a genome containing many genes that are thought to have correlated activity and disease potential. The Roche methodology is applied to different mouse strains to produce collective identification of the regions of interest. In quite a similar fashion, the Connectron methodology has been applied to many different bacterial, Archeal and eukaryotic genomes. Some of the genomes in the public domain are strain-like variations of each other.

Applying a computational methodology to a variety of genomes and now transcriptomes is now an accepted approach to understanding how cells and tissues function and how disease arises and may be eventually remedied.

Addendum 3

Helen Pearson in a News Feature in the November 16th, 2006 issue of *Nature* provides a discussion of the possibility that there are many different codes in DNA. Within this article Dr. Jussi Taipale at the University of Helsinki in Finland argues "the biggest obstacle after the sequencing of the genome has been to understand how genes are regulated and how we can see that from the sequence". (This inter alia is what the Connectron methodology is trying to do). Taipale goes on to say: " it's a more complex code than the genetic code." The Connectron methodology provides a sequence-based approach to trying to understand how gene expression is regulated.

Pearson argues: "A human cell has to fit about two meters of DNA into a nucleus a few micrometers in diameter; that requires packing into together with proteins in a complex hierarchy of folding back and wrapping around. The fundamental element underlying all this packaging is the nucleosome – 147 base pairs of DNA wrapped about a globule of eight proteins called histones." Pearson goes on to say: " It has been known for more than two decades that in the test-tube certain sequences are more likely to be packaged up in nucleosomes. But in the real hustle and bustle of the cell, it was unclear to what extent such preferences get honored." Pearson mentions, "Dr. Eran Segal at the Weizmann Institute in Rehovot, Israel and his colleagues came the closest yet to defining a code for the position of the nucleosomes." Segal and his colleagues have tried to define this code with only a database of 377 nucleosomes. Typical Connectron computations use the whole genome or transcriptome. There may be some sequence-based pre-disposition to binding nucleosomes. Computationally identified Connectron sequences (i.e. the T1s – the left-flanking sequences - and the T2s – the right-flanking sequences) might also be another level of code. Pearson goes on to say: "DNA seems well adapted to supporting a number of codes."

Pearson then discusses the controversy of the meaning of long-range patterns in DNA – whether these patterns are biologically meaningful or not. Pearson closes the article with a quote from Wyeth Wasserman at the University of British Columbia in Vancouver who says: " Computer scientists think they can just walk in the door and solve things. But they come to realize you need biology too." This is the heart of the discussion between the Applicant and the Examiner. The question is whether the Connectron methodology and it application to various genomes and transcriptomes contributes to this process of developing insights and understanding of biological systems. The computational results from the mouse transcriptome clearly demonstrate that finding the Connectron patterns and then analyzing them produces higher-level non-random patterns. We have shown that the application of the Connectron methodology per se has scientific and intellectual utility even though the specific mechanism of gene expression regulation may still not be resolved. Just as clearly, the Connectron methodology is an invention that produces concrete results.

Addendum 4

Davidson and Carver in an announcement from the University of Iowa dated November 13th, 2006 have shown that microRNAs are produced from the 'Junk DNA' regions of genomes and that the molecular machinery used to produce these microRNAs is different from that used to produce RNA for protein translation.

The mouse transcriptome data from the RIKEN includes short RNA transcripts that are produced from DNA in the introns of proteins. The characteristics of these microRNAs is that they different lengths but that all the transcripts have a common left boundary for positive-strand transcription and a common right boundary for negative-strand transcription. The Connectron methodology when applied to the mouse transcriptome has shown that Connectrons arise from these microRNAs thus leading to the expectation that these microRNAs control the expression of genes and other non-coding events.

It is thought that local under-coiling just to the left of the start of transcription allows an unguarded polymerase to begin transcription. In order to conserve global neutrality of super-coiling, there must be a region of over-coiling somewhere to the right of the start of transcription – for positive-strand events. There is no sharp termination of transcription signal as there is for protein transcription but rather the polymerase runs into the region of over-coiling and statistically stops transcription.

While molecular biologists stir and poke with their various methodologies, the computational methodologies can make clear predictions. For example, the Connectron predictions of Gene-Coding and Non-Coding transcription regulation will reduce physical experimentation by many orders of magnitude. The molecular biologists are looking for a theory. Since the goal of inventions within the culture, in general, is to increase efficiency and to stimulate new ways of doing business, the Applicant believes that the Connectron invention is entitled to protection because it has already shown that it can produce scientific utility (i.e. it makes predictions that can be validated by physical experiments).

Addendum 5

Dr. Isidore Rigoutsos and colleagues at the IBM Watson Research Center at Yorktown Heights have shown in a paper published in the PNAS on April 25th, 2006 that using an unsupervised pattern identification process they have discovered in the human genome multiple copies of variable-length patterns that occur more frequently that would be expected by chance. Rigoutsos et al call these patterns "Pyknons". Looking at the reverse complementing properties of the RNA that would be produced from the Pyknon sequences, Rigoutsos says that these sequences will "form double-stranded, energetically stable, hairpin-shaped RNA secondary structure". The Pyknon sequences are typically 60-80 bases in length – about the length of tRNAs.

Rigoutsos goes on to say: " These unexpected findings suggest potential unique functional connections between the coding and non-coding parts of the human genome.

The Applicant in the draft paper supplied for the record at the interview has studied (since the initial development of the Connectron methodology) the mouse transcriptome using data supplied by the Japanese National Genome Project (RIKEN) in Yokohama. The Connectron methodology applied to this transcriptome produces in an unsupervised manner, multiple instances of conserved copies of patterns that occur above chance expectation levels.

Addendum 6

Dr. Tamara Frazier at Stanford University in Palo Alto discusses the application of computational methods to describe and establish the utility of DNA sequences for the purposes of patenting. Much Frazier's discussion is devoted to the history of EST sequence patenting. Although in the present basic methods patent application under consideration we have been forced by PTO protocol to identify and document the DNA sequences in our many examples of Connectrons, nowhere in the application are we claiming these sequences perse. These DNA sequences are merely used as examples of the four-sequence Connectron relationships.

The Frazier review is useful because it helps to show how the focus of science has shifted in ten years from patenting sequences to patenting methodologies that show relationships between sequences. Whereas in the EST time, the focus was on showing the use of a gene, in the present time, the focus has shifted to understanding global correlations between a huge variety of Gene-Coding and Non-Coding transcription events.

For example, Frazier does not discuss SNPs at all. Scientists were just beginning to realize the importance of SNPs in the EST days. When SNPs occur in Gene-Coding and Non-Coding regions of the genome, they can produce changes in the Connectron control of transcription.

The computational methodology (as outlined in the hierarchy of flow diagrams) finds instances of the four-sequence Connectron relationship. For a given genome or transcriptome, the Connectron methodology produces a set of predictions as to which transcription events will control other transcription events. The original Application talks in terms of gene expression regulation but the mouse transcriptome work has shown that both Gene-Coding DNAs and Non-Coding DNAs both produce transcription regulation.

The USPTO Appeal Board has argued that the utility of the Connectron methodology (i.e. the production of a set of predictions of transcription regulation) is independent of the process of validating the predictions.

The Frazier review helps us to realize that our understanding of what is interesting and important in science changes (very rapidly) in time. No serious person argues about EST

patenting toady. That issue is settled. The power of computation has vastly increased as we have gone in ten years from giga-op processing levels to tera-op processing levels and soon to peta-op processing levels. Discussion about the function of single a gene that was interesting ten years ago is essentially no longer of interest. The utility today of computation is in theory formation from genomic/transcriptomic data! Today's already high levels of computation allow us to extract and understand the coherence that exists in genomes and transcriptomes.

Addendum 7

R. Thenmalarchelvi a doctoral student of Dr. N. Yathindra at the University of Madras in Chennai has written a thesis on the formation of sequence-dependent RNA-DNA-DNA triple-strand Hoogsteen helices. Molecular mechanics was used to study the binding RNA in the major groove of the DNA double-strand helix. The Contents and Preface of this thesis is presented. The thesis is very technical. A shorter scientific paper is forthcoming and will hopefully resolve the question of the stability of generalized sequence-dependent triple-stranded helices through concise generalizations.

The classical Hoogsteen triple-helices have a restricted range of sequences. Thenmalarchelvi states: "One of the major outcomes of this study is that the residual twist may be responsible for sequence dependent non-uniform structural variations in DNA triplexes comprising non-isomorphic base triplets. As with the binding of Zinc-finger DNA binding proteins, the strong RNA to double-strand DNA bindings are by means of hydrogen bonds whereas the weaker base bindings are mediated by hydrophobic contacts and water molecules.

ADDENDUM #1

Computational Analysis as a Mode of Biological Discovery

The past ten years have witnessed a dramatic, paradigm-shifting transformation in biology. The elucidation of the complete human genome sequence, plus that of primates, and an increasing number of mammals, including "mouse," *Mus musculus*, "rat" *Rattus norvegicus*, and a host of others has provided an immense wealth of data. The word immense is particularly relevant since it is self evident that the sheer amount of data in these genomes, or indeed even in a "simple" bacterial genome, cannot be analyzed without sophisticated computational tools. But what is the goal of this analysis? That is to say, what major questions remain unresolved and how can they be studied by computational analysis. It is worth noting, for instance, that while only about 5% of the mouse genome codes for proteins, nearly 80% of it is transcribed to RNA molecules. While the functions of some of these non-protein coding RNAs is clear, the functrion of the vast majority of this RNA is completely unknown. How can one probe this data, the mouse transcriptome, for the potential for biological function? The German philosopher, Immanual Kant succinctly captured this state of affairs,

Concepts without observations are empty, observations without concepts are blind...Only through their union can knowledge arise.

Kant, I. *Critique of Pure Reason*, University of Virginia Library, Electronic Text Center, Topic I, Part II, 45

In the biological sciences, computational methods that simulate biological behavior are gaining increasing performance.

"The massive acquisition of data in molecular and cellular biology has led to the renaissance of an old topic: simulations of biological systems. Simulations, increasingly paired with experiments, are being successfully and routinely used by computational biologists to understand and predict the quantitative behaviour of complex systems, and to drive new experiments. Nevertheless, many experimentalists still consider simulations an esoteric discipline only for initiates. Suspicion towards simulations should dissipate as the limitations and advantages of their application are better appreciated, opening the door to their permanent adoption in everyday research.[1]"

In fact, the complexity of biological systems means that approaches that do not exploit computational models are likely to have a very difficult time designing critical experiments, and given the cost of biological research, computational modeling can save a significant amount of time and money.

---

[1] Barbara Di Ventura, Caroline Lemerle, Konstantinos Michalodimitrakis, and Luis Serrano, From *in vivo* to *in silico* biology and back, *Nature* **443**, 527-533 (5 October 2006) | doi:10.1038/nature05127

Computational biology will play a critical role in analyzing massive amounts of genomic information. The role of computational biology in answering critical questions about the mechanism by which gene expression is regulated is summarized below:

Computational methods have become intrinsic to modern biological research, and their importance can only increase as large-scale methods for data generation become more prominent, as the amount and complexity of the data increase, and as the questions being addressed become more sophisticated. All future biomedical research will integrate computational and experimental components. New computational capabilities will enable the generation of hypotheses and stimulate the development of experimental approaches to test them. The resulting experimental data will, in turn, be used to generate more refined models that will improve overall understanding and increase opportunities for application to disease. The areas of computational biology critical to the future of genomics research include:

- New approaches to solving problems, such as the identification of different features in a DNA sequence, *the analysis of gene expression and regulation,* the elucidation of protein structure and protein_protein interactions, the determination of the relationship between genotype and phenotype, and the identification of the patterns of genetic variation in populations and the processes that produced those patterns
- Reusable software modules to facilitate interoperability
- Methods to elucidate the effects of environmental (non-genetic) factors and of gene_environment interactions on health and disease
- New ontologies to describe different data types
- Improved database technologies to facilitate the integration and visualization of different data types, for example, information about pathways, protein structure, gene variation, chemical inhibition and clinical information/phenotypes
- Improved knowledge management systems and the standardization of data sets to allow the coalescence of knowledge across disciplines.[2]

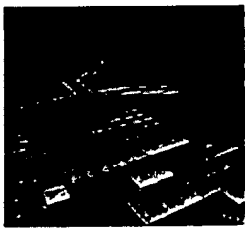
In sum, "advances in our understanding of genomic sciences and the development of new and more robust tools to investigate and analyze biological systems has led to an emphasis on analyzing biological systems at multiple levels. Thus, there is a need to integrate different types of data into a comprehensive 'systems' view. These include improved analytical and visualization tools and the ability to integrate different types of data into a comprehensive view of biological processes. These approaches are beginning to provide new and profound insights into human biology with the potential for new effective interventions in treating and preventing human diseases.[3]"

---

[2] Francis S. Collins, Eric D. Green, Alan E. Guttmacher and Mark S. Guyer, A vision for the future of genomics research, Nature **422**, *835-847 (24 April 2003).*

[3] Jeffrey M. Trent, Andreas D. Baxevanis, Chipping away at genomic medicine, *Nature Genetics* **32**, 462 - 462

ADDENDUM #2



A new computational method for rapid, precise analysis of genetic variations

Basel, 22 October 2004

New Genomic Method Can Identify Disease-Causing Genes with Unprecedented Precision and Speed

A novel computational method to detect disease-causing genes accurately and rapidly was announced by Roche scientists in the October 22 issue of *Science*. This approach, another innovation in computational genetic analysis from Roche scientists, promises to accelerate markedly the discovery of mouse correlates of genetic risk factors for human disease. The new approach enables researchers to identify a single causative genetic factor by correlating a pattern of observable physiological or pathological differences among selected strains of mice with a pattern of genomic variation. Using conventional methods, pin-pointing a gene contributing to disease risk could take five scientists five years. With Roche's latest innovation, which has up to 1,000-fold greater precision than current methods, a single researcher may accomplish the task in a single afternoon. The method takes advantage of the block-like patterns of genomic variation in selected mouse strains, as illustrated on the cover of *Science* in which the article appears.

"Our hope is that this new computational approach will increase the utility of the vast amount of DNA sequence information available today and help researchers more fully leverage mouse models of human disease to identify genes contributing to disease risk and drug response," said Gary Peltz, M.D., Ph.D., head of Genetics and Genomics at Roche Palo Alto. "It will help researchers understand the relationship between trait differences and variations in the mouse genome, which will move us a long way toward understanding the impact of human genetic differences. As that happens, we should be able to translate genetic data more effectively and efficiently into the development of both novel diagnostic tools and new medicines to treat human diseases."

In this regard, Roche Palo Alto is engaged in research with several leading universities and government institutions to leverage the power of the new computational technique. The studies are directed toward better understanding the

genetic causes of a range of human diseases and toward pharmacogenetic analysis of how various drugs that are used commonly to treat disease work in humans.

The paper, entitled "***In Silico* Genetics: Identification of a Novel Functional Element Regulating H2-Eα Gene Expression,**" reports that the new computational algorithm correctly identified the genetic basis for strain-specific differences in several biologically important traits, including differences in drug metabolism. The examples presented in the paper demonstrate the ability of the methodology to identify causative genetic factors accurately for a wide range of trait data. The technique also has the potential to uncover currently unknown genetic factors contributing to a host of different diseases.

Roche scientists first published a computational method for mouse genome analysis in the June 8, 2001 issue of *Science*. That method predicted regions of a mouse chromosome responsible for a trait difference. The predicted regions contained hundreds of genes and the results were assessed by relative (percentile ranking) statistical criteria. The new method offers the same analytic speed, but is much more exact, linking a single gene to a trait difference. This method eliminates the need for follow-up studies to mine large chromosomal regions, saving researchers from months to years of experimentation. In addition, the results are assessed by absolute (p-value) statistical criteria, which give researchers greater confidence in their analyses.

The pattern of genetic variation analyzed by this new computational method was created by mining a database of common genetic markers, called single nucleotide polymorphisms (SNPs), covering 1,900 genes across 16 commonly used inbred mouse strains. That database was created by Roche scientists in Palo Alto, Alameda, Calif., Basel, Switzerland, and was partially sponsored by a National Human Genome Research Institute Grant. It was recently selected as the top SNP database by respondents to a survey of scientists conducted by *Genome Technology and GenomeWeb Daily News*. The genetic pattern maps are now available to the public for the first time as part of the Roche SNP database web site. The web site delivers a wealth of genetic information about many mouse strains that are commonly used to model human disease.

Because the mouse genome is similar to that of humans, the mouse is the most commonly used experimental model for studying human disease, and the "mouse to man" approach is widely used. Since analyses of mouse genetic models by traditional methods are very time-consuming and costly, this novel computational approach represents a major advance for this entire field of research.

Study participants from Roche included Guochun Liao, Jianmei Wang, Jingshu Guo, John Allard, Janet Cheng,

Anh Nguyen, Gary Peltz, and Jonathan Usuka from the Roche Palo Alto campus, and Dorothee Foernzler from the Roche Center for Medical Genomics in Basel, Switzerland. Other study participants included: Steve Shafer from Stanford University, Stanford, California; Anne Peuch from the Centre National de Génotypage, France and John D. McPherson from the Washington University School of Medicine, St. Louis, Missouri.

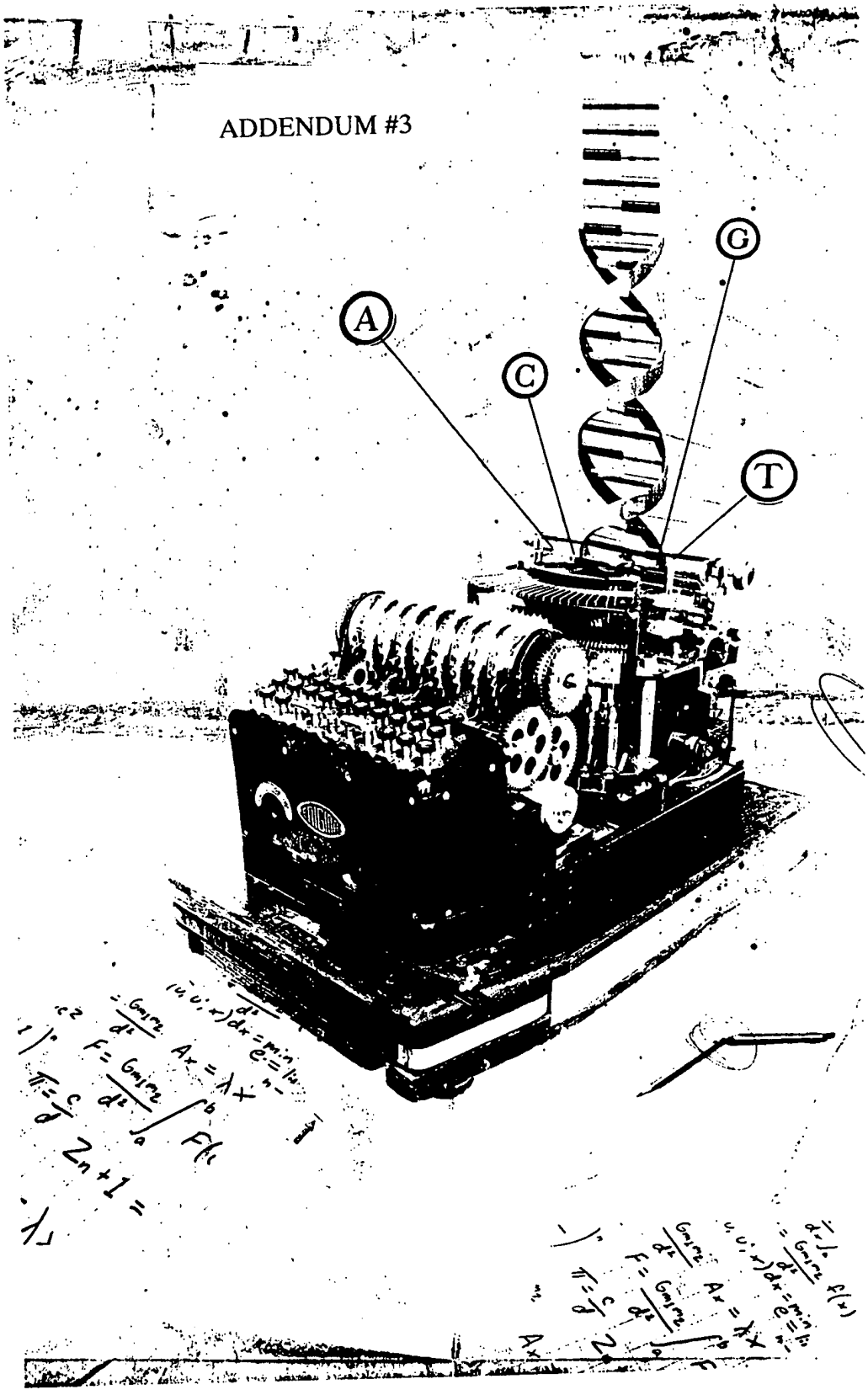
About Roche

Headquartered in Basel, Switzerland, Roche is one of the world's leading research-intensive healthcare groups. Its core businesses are pharmaceuticals and diagnostics. As a supplier of innovative products and services for the prevention, diagnosis and treatment of disease, the Group contributes on a broad range of fronts to improving people's health and quality of life. Roche is number one in the global diagnostics market, the leading supplier of medicines for cancer and transplantation and a market leader in virology. Roche employs roughly 65,000 people in 150 countries and has R&D agreements and strategic alliances with numerous partners, including majority ownership interests in Genentech and Chugai.

Further information:

- Genes and Health:
http://www.roche.com/pages/facets/22/gene2_e.pdf

print this page No date available



US NATL CRYPTOL MUS./D. ALLISON

CODES AND ENIGMAS

There's more than one way to read a stretch of DNA, finds **Helen Pearson** — and we need to understand them all.

"Itwasknownthattheywerealittleacquaintedbutnotasyllableofrealinformationcouldemmaprocureastowhathetrulywas..." Reduce it to just a sequence of letters, and even a delicate phrase from Jane Austen's *Emma* becomes virtually impenetrable gobbledygook. So it was something of a triumph for Simon Shepherd when, in 2001, an algorithm he had written reconstructed all of *Emma*, word for separated word, from just such an uninterrupted string, despite being unacquainted with English vocabulary or syntax. The software worked out which groupings of letters were most likely to appear together, and thus have distinct meanings.

Shepherd, a researcher at the University of Bradford, UK, picked up much of his expertise during ten years cracking Russian codes in British Naval Intelligence. But he was not really interested in *Emma* — that was just a demonstration. His real goal was the far longer sequences of As, Gs, Cs and Ts that make up the world's genomes. Within those strings there is information that no one knows how to extract — codes that regulate, control or describe all sorts of cellular processes. And if the information is there, Shepherd thinks that number crunching should be able to pry it loose. "We are treating DNA as we used to treat problems in intelligence," he says. "We want to break the code at the most fundamental level."

That DNA contained at least one code was realized as soon as the molecule's structure was discovered. That code, cracked in the 1950s and 1960s, parses passages of DNA into three-letter combinations that correspond to particular amino acids. This is a code in the strictest sense; input determines output.

But researchers now know that there are numerous other layers of biological information in DNA, interspersed between, or superimposed on, the passages written in the triplet code. Human DNA contains tissue-specific information that instructs brain or muscle cells to produce the suite of proteins that make them brain or muscle cells. Other signals in the sequence help decide at what points DNA should coil around its scaffolds of structural proteins. These are the codes that computer buffs such as Shepherd want to crack with raw processing power — and that mainstream biologists are attacking, too, although using a rather more lab-based approach. "We need all these codes together to understand the dynamics of the cell," says computational biologist Manolis Kellis at the Massachusetts Institute of Technology in Cambridge.

The DNA sequence contains information

not just about the make-up of proteins but also about the interactions of DNA with some of those proteins, and the diverse antics of RNA. The analysis of DNA sequences is revealing patterns that have meanings at all of these levels. "Biology has probably figured out a way to squeeze every bit of information from that molecule it can," says Jason Lieb, who studies DNA–protein interactions at the University of North Carolina at Chapel Hill.

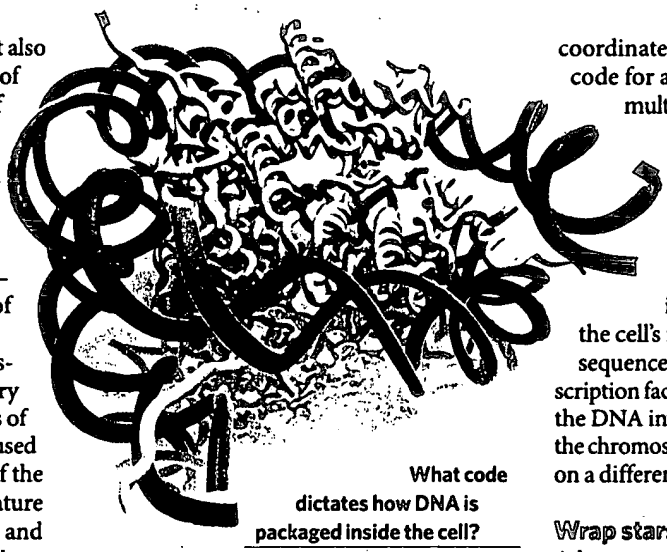
The code that is currently most exercising the minds of geneticists is the 'regulatory code' that directs the production of suites of proteins tailored to specific cell types and used at specific times. The idea is that many of the genes switched on in DNA contain signature sequences in 'promoter' regions nearby and 'enhancer' regions that may be millions of base pairs away. In a blood cell, say, these signature sequences might be bound by proteins A, B, C and D, whereas genes switched on in skin may be regulated by signature sequences that bind proteins B, C, Y and Z.

"The biggest obstacle after the sequencing of the genome has been to understand how genes are regulated and how we can see that from the sequence," says Jussi Taipale, who studies gene regulation at the University of Helsinki, Finland. "It's a more complex code than the genetic code." The first difficulty is the sheer scale of the problem. Human cells contain more than 20,000 protein-coding genes, roughly 1,500–2,000 transcription factors, which switch genes on and off, and numerous other regulatory proteins and RNAs that direct their production. The possible permutations and combinations are bewildering.

Lost in translation

One way to start solving the regulatory code en masse would be to find all the positions where each of the regulatory proteins binds within the genome. Many transcription factors show a penchant for binding specific short motifs in DNA, such as a six-letter sequence. In theory, a computer could scan for any such motifs that occur more often than might be expected by chance.

But there are drawbacks. For one thing, a given six-base-pair sequence will sometimes be a binding site and sometimes not, probably depending, in part, on whether the DNA is folded up in a way that prevents transcription factors from gaining access. For another, the way that these sites are recognized is not as specific as the binding between the bases that translate the triplet code into protein. Transcription factors recognize DNA sequences from the effects of the sequence on the outside of the helix, and although this recognition is still sequence dependent, it is not quite so precise. Some of these proteins will bind to a range of related sequences — sometimes more tightly, sometimes less so — and those subtleties of affinity, like the nuances of a social embrace, may themselves have biological meaning.



PHANTATOMIX/SPL

What code dictates how DNA is packaged inside the cell?

Len Pennacchio at the Lawrence Berkeley National Laboratory in California and his colleagues have begun to fathom some of these subtleties by identifying a rudimentary tissue-specific code for the human brain[1]. They teased out the relevant enhancers from the human genome by comparing the human sequence to those of distant relatives such as the pufferfish (*Takifugu rubripes*), pulling out regions that didn't describe proteins but that evolution had nevertheless deemed important enough to keep intact. They then systematically inserted 167 such regions into mouse embryos and found that 45% of them provided tissue-specific ways to switch on genes.

The team identified four enhancers that boost gene activity in the developing forebrain and share several short-sequence motifs that are presumably binding sites for control proteins. By searching for similar signature sequences in the human genome, they located other forebrain enhancers, suggesting that they have found some of the sequence information that 'means' brain-specific in the regulatory code.

"We are treating DNA as we used to treat problems in intelligence." — Simon Shepherd

Taking a slightly different tack, Richard Young at the Whitehead Institute for Biomedical Studies in Cambridge, Massachusetts, and his colleagues have come up with a preliminary code that distinguishes human embryonic stem cells[2]. They extracted human DNA bound by three key transcription factors and determined all the sequences to which those proteins chose to bind. The proteins recognize sequences near genes that need to remain active for stem cells to stay stem cells; they also recognize other sites where they seem to help shut down the genes needed for the stem cells to differentiate into other cell types. So these proteins, in combination with others, seem to stop stem cells from becoming other cell types.

Many researchers are now talking about a coordinated effort to identify the regulatory code for all human transcription factors in multiple tissues. But they are unlikely to resolve this code without simultaneously extracting other layers of overlapping information in DNA.

A section of DNA can contain two or more layers of information that are used at different times or in different ways depending on the cell's requirements. So whether a given sequence is read as a binding site for a transcription factor to some extent depends on how the DNA involved is packaged at that point in the chromosome — and that packaging depends on a different code stored in the DNA.

Wrap stars

A human cell has to fit about two metres of DNA into a nucleus a few micrometres in diameter; that requires packing it together with proteins in a complex hierarchy of folding back and wrapping round. The fundamental element underlying all this packaging is the nucleosome — 147 base pairs of DNA wrapped around a globule of eight proteins called histones. Up to 90% of DNA is bundled up into nucleosomes, and their position influences the DNA's activity. Sequences wrapped up in nucleosomes are often less accessible to transcription factors and so less likely to be transcribed. It has been known for more than two decades that in the test-tube certain sequences are more likely to be packaged up in nucleosomes. But in the real hustle and bustle of the cell, it was unclear to what extent such preferences get honoured.

Earlier this year, Eran Segal at the Weizmann Institute of Science in Rehovot, Israel, Jonathan Widom at Northwestern University in Evanston, Illinois, and their colleagues came the closest yet to defining a code for the position of nucleosomes[3]. They took DNA wrapped up in nearly 200 yeast nucleosomes, and 177 from chickens, and exposed it to enzymes that would eat up all sequences in between the nucleosomes. They then sequenced the DNA left intact in the nucleosomes, and used computational methods to align the sequences and search for common patterns.

The team came up with a set of rules that could predict where more than 50% of nucleosomes lie in yeast and chicken DNA. "It's much less than perfect but way better than random," Widom says. The main rule is that the sequences AA, TT or TA are more likely to be found where the spiralling DNA backbone grazes the histone — they seem to help the DNA bend around the protein core.

But Segal and Widom's rules can't predict the position of a significant fraction of the

nucleosomes. DNA's overlapping codes mean that an individual nucleosome might be usurped if regulatory proteins are already tightly bound there. The nucleosome code depends on the regulatory code, just as the regulatory code depends on the nucleosome code. In addition, the position of a nucleosome might be influenced by the way in which the nucleosome-wrapped sequence is folded and condensed yet further. "The code specifies the initial state and the cell can mess with what happens afterwards," says Oliver Rando, who studies nucleosome positioning at Harvard University.

"It's much less than perfect but way better than random."
— Jonathan Widom

The goal now is to find codes that govern those larger-scale features of DNA packaging, such as how the nucleosomes are twisted up into a cable of chromatin and eventually coiled into the tightly interwoven ropes of the chromosome. As yet, though, researchers have not found landmarks equivalent to nucleosomes that can guide the search for meaning — nor is it clear that they will. "There could be diffuse information spaced at hundreds of kilobases that helps package even larger pieces of the genome together," says Lieb. "Or it could be that the exact position of those structures is not important."

Room for manoeuvre

DNA seems well adapted to supporting a number of codes. For a start, only 1–2% of the human genome is occupied with protein-coding sequences, which leaves plenty of intervening DNA to hold other information. But many stretches of DNA in humans and other organisms manage to multitask: a sequence can code for a protein and still manage to guide the position of a nucleosome. This is possible because the triplet code is 'degenerate'. Several slightly different triplets can code for the same amino acid, and many positions in a protein can be filled by different amino acids — so different sequences can effectively mean the same thing. This allows other signals to be imprinted on top of the first — especially when those other signals are themselves encoded with some slack.

This elegance is surely the handiwork of evolution — and if the way in which that hand had worked to solve these problems were clearer, the simultaneous decoding of all the messages involved might become easier. Perhaps ancestral organisms had simpler sequence patterns that evolution has optimized, taking advantage of its degeneracy to layer in additional information that helped organisms acquire extra complexity. Hanspeter Herzel, who specializes in statistical analyses of DNA at Humboldt University, Berlin, speculates that the space constraints of the cell may have favoured the development of nucleosomes that wound up unruly DNA — and that their existence then encouraged the evolution of a nucleosome code in the sequence because this lowered the energetic cost of coiling up DNA. But as yet such ideas, and any help they might offer, remain tentative. "We don't really have a phylogeny of these signals," he says.

And in some cases, it seems that evolution may have generated patterns that have no clear biological function. In 1992, Gene Stanley at Boston University, Massachusetts, and his co-workers created waves when they suggested that there were patterns in DNA that spanned hundreds and thousands of base pairs[4]. Stanley used the types of statistical techniques that identify correlations in climate and financial data and applied them to all the DNA sequences available in databases at the time.

Essentially, the study showed that a region with a particular chemical composition, such as one loaded with the bases A and G, is likely to be followed by a similar region hundreds or thousands of base pairs away, and that the probability of this pattern declines in a predictable way with distance. It also found that this correlation existed predominantly in DNA that did not code for protein, leading Stanley to propose that DNA previously written off as junk actually carries biological information.

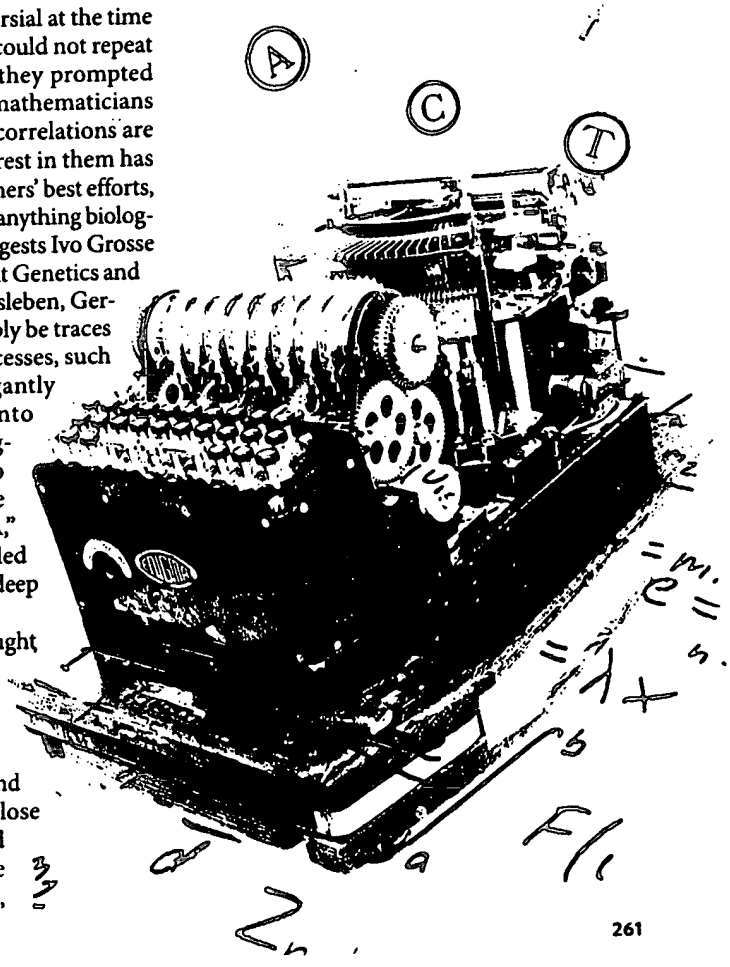
The findings were controversial at the time because several other groups could not repeat aspects of the analysis, and they prompted huge interest in DNA from mathematicians and physicists. Today, these correlations are thought to be real — but interest in them has faded because, despite researchers' best efforts, the patterns have not revealed anything biologically important. Perhaps, suggests Ivo Grosse of the Leibniz Institute of Plant Genetics and Crop Plant Research in Gatersleben, Germany, the patterns could simply be traces of random evolutionary processes, such as the erosion patterns elegantly but accidentally carved into sandstone by the wind. "Long-range correlations definitely do exist, but I don't think it's some supercode imprinted in DNA," Grosse says. "We just stumbled on a feature with probably no deep biological meaning."

But to some people the thought of order with no meaning is an affront. To such minds, the idea of teasing out nature's secrets with little more than mathematical cunning and processing power will never lose its allure. When Shepherd and his graduate student Natalie Kay, in unpublished work, ran the software that they had tried out on *Emma* over the (admittedly small) genome of Ebola virus, it identified as meaningful some sequences that, at the time, bore no annotations in genetic databases. Only later, Shepherd says, were these motifs recognized by biologists as passages that control the activity of genes or mark their ends. He thinks that approaches based on almost pure number crunching will go on to rock the field: "I firmly believe that major advances in this over the next 20, 30, 50 years will be made by the theorists, not the medics."

But researchers versed in the complexities of how DNA and proteins actually work remain convinced that their type of knowledge will remain vital to sorting the meaningful from the circumstantial. When the triplet code was first being studied, there were any number of fanciful mathematical and logical approaches to it — but the approaches that paid off were the ones informed by the greatest degree of biological insight. "Computer scientists think they can just walk in the door and solve things," says bioinformatics expert Wyeth Wasserman at the University of British Columbia in Vancouver, Canada. "But they come to realize you need biology too." □

Helen Pearson is a reporter for *Nature* based in New York.

1. Pennacchio, L. A. *et al. Nature* doi:10.1038/nature05295 (2006).
2. Boyer, L. A. *Cell* **122**, 947–956 (2005).
3. Segal, E. *et al. Nature* **442**, 772–778 (2006).
4. Peng, C.-K. *et al. Nature* **356**, 168–170 (1992).



US NATL CRYPTOL. MUS./D. ALLISON

X-Originating-IP: [207.172.157.102]
X-IronPort-AV: i="4.09,426,1157342400";
d="scan'208,147"; a="348328867:sNHT51270344"
To: "Richard J. Feldmann" <rjfeldma@erols.com>
Subject: University of Iowa Scientists Explore Function of 'Junk DNA' ADDENDUM #4
From: lyle.middendorf@licor.com
Date: Wed, 15 Nov 2006 17:14:37 -0600
X-Junkmail-Status: score=10/50, host=mr14.lnh.mail.rcn.net
X-Junkmail-SD-Raw: score=unknown,
refid=str=0001.0A090206.455B9EBE.00BC,ss=1,fgs=0,
ip=63.151.144.205,
so=2006-05-09 23:27:51,
dmn=5.2.121/2006-09-27
X-Mirapoint-Loop-Id: f958af4dea3b808f65ab570e60a5ce2d

University of Iowa Scientists Explore Function of 'Junk DNA' 11/13/06 -- University of Iowa scientists have made a discovery that broadens understanding of a rapidly developing area of biology known as functional genomics and sheds more light on the mysterious, so-called "junk DNA" that makes up the majority of the human genome.

The team, led by Beverly Davidson, Ph.D., a Roy J. Carver Biomedical Research Chair in Internal Medicine and UI professor of internal medicine, physiology and biophysics, and neurology, have discovered a new mechanism for the expression of microRNAs -- short segments of RNA that do not give rise to a protein, but do play a role in regulating protein production. In their study, Davidson and colleagues not only discovered that microRNAs could be expressed in a different way than previously known, they also found that some of the junk DNA is not junk at all, but instead consists of sequences that can generate microRNAs.

Davidson and her colleagues, including Glen Borchert, a graduate student in her lab, investigated how a set of microRNAs in the human genome is turned on, or expressed. In contrast to original assertions, they discovered that the molecular machinery used to express these microRNAs is different than that used to express RNA that encodes proteins. Expression of the microRNAs required an enzyme called RNA Polymerase III (Pol III) rather than the RNA Polymerase II (Pol II), which mediates expression of RNA that encode proteins. The study is published in Nature Structural and Molecular Biology Advance Online Publication (AOP) on Nov. 12.

"MicroRNAs are being shown to play roles in cancer and in normal development, so learning how these microRNAs are expressed may give us insight into these critical biological processes," said Borchert, who is lead author of the study. "Up to now it's been understood that one enzyme controls their expression, and we now show that in some cases it's a completely different one."

Genes that code for proteins make up only a tiny fraction of the human genome. The function of the remaining non-coding sequence is just beginning to be unraveled. In fact, until very recently, much of the non-coding sequence was dismissed as junk DNA. In 1998, scientists discovered that some DNA produced small pieces of non-coding RNA that could turn off, or silence, genes. This discovery won Andrew Fire and Craig Mello the 2006 Nobel Prize for medicine or physiology. Since their discovery, the field has exploded and small, non-coding RNAs have been shown to play an important role in development and disease in ways that scientists are only just beginning to understand.

"Not so many years ago our understanding was that DNA was transcribed to RNA, which was then translated to protein. Now we know that the levels of control are much more varied and that many RNAs

don't make protein, but instead regulate the expression of proteins," Davidson explained. "Non-coding RNA like microRNAs represent a set of refined control switches, and understanding how microRNAs work and how they are themselves controlled is likely to be very important in many areas of biology and medicine."

Over 450 microRNAs have been identified in the human genome. Learning how they are turned on and in what cells and what they do, may allow scientists to turn that knowledge to their advantage as a medical tool.

Source: University of Iowa

The Use of Computational Methods to Describe and Establish Utility of a DNA Sequence for Purposes of Patenting

Tamara Fraizer

I. INTRODUCTION

Computers and the internet have, in the matter of a few decades, changed the nature of personal communication, business, and scientific research. The creation of large gene and protein databases and the development of sophisticated methods for analyzing sequence data via the web have, for example, transformed certain aspects of molecular biology and genetics into the information sciences now known as genomics and bioinformatics. Indeed, the typical research biologist now combines work at the bench with work online, and knows both chemistry and computational methods. Meanwhile, companies such as Incyte and Celera are specializing in the production and analysis of genetic information, leaving other companies to pursue the development of particular pharmaceutical products.[1]

The recent changes in methods of biological research and business create significant challenges for the definition and defense of intellectual property rights relating to genetic research.[2] Our legal system, an institution of resilience rather than reform, is adapting to the new world. Together, the Court of Appeals for the Federal Circuit[3] (CAFC) and the United States Patent and Trademark Office (USPTO) are establishing the precedents and procedures needed to assess whether and how particular genetic discoveries can be patented. The process is slow and imperfect, though, and the pace of scientific advancement has made many of the CAFC's rulings appear inadequate if not obsolete. Nonetheless, the USPTO has responded in timely and pertinent ways, interpreting the CAFC's rulings in guidelines that are used by examiners in evaluating patent applications.

The CAFC and the USPTO are struggling most notably to adapt precedents and procedures to a fundamentally new type of invention: myriad isolated cDNA sequences whose functions are inferred from computational analysis of existing annotated databases of genetic sequences. Many people have argued that such inventions are merely "information about the natural world" and therefore should

---

[1] Randall Scott, President and Chief Scientific Officer of Incyte Genomics, described the new pharmaceutical industry as being vertically rather than horizontally integrated. Thus, instead of one company conducting all aspects of research and development, some companies provide the data needed for early stages of R&D while other companies direct the commercial development of particular products. Randall Scott, President and Chief Scientific Officer, Incyte Genomics. Prepared Statement at Hearing Before the House of Representatives Subcommittee on Courts and Intellectual Property of the Committee on the Judiciary, on Gene Patents And Other Genomic Inventions. 106th Congress, 2nd Session, July 13, 2000 [hereinafter Congressional Hearing on Genomic Inventions]. Interestingly, the agricultural industry appears to remain horizontally integrated, with most aspects of the industry dominated by companies such as DuPont and Pioneer.

[2] The National Academy of Science has acknowledged the significance of these challenges and is conducting a two-phased project study on "Intellectual Property Rights in the Knowledge-Based Economy". *See* http://www4.nas.edu/cp.nsf.

[3] The CAFC was created in 1982 as a speciality court that would hear appeals from all the federal district courts involving patent issues. Many of the CAFC judges have technical backgrounds and all are more familiar with patent issues than the typical court of appeals judge. Thus, the creation of the CAFC has helped to create a systematic and sensible body of patent law.

not be patentable.[4] Participants at the 1996 International Strategy Meeting on Human Genome Sequencing endorsed the idea that "all genomic DNA sequence information should be "freely available and in the public domain in order to encourage research and development and to maximize its benefit to society".[5]

The ease with which researchers can now obtain cDNA sequences of unknown function, and compare them to sequences of known functions, stands in contrast to the state of the art in the 1980s, when researchers worked diligently to determine the actual nucleotide sequence for proteins of known function. These contrasting states of the science have raised two legal issues: (1) whether the invention, i.e. nucleotide sequence, is possessed by the inventor and adequately described, and (2) whether the invention, i.e. cDNA fragment, has a real world utility.

The issue associated with the earlier situation—i.e. patents claiming an unknown (but knowable) sequence of experimentally known function—has been addressed by the CAFC. In the early 1990s, the CAFC chose to assess DNA as it would any chemical compound. To claim a chemical compound as a composition of matter, the inventor must describe the compound's structure. Therefore, the court found that describing a protein's function and a method for isolating its DNA was not enough to claim the gene. Rather, the inventor had to describe the DNA, which was most obviously done by giving its nucleotide sequence. Recently, in January 2001, the USPTO published guidelines for assessing the adequacy of the description of inventions, consistent with the CAFC's decisions, and applied them to contemporary scientific scenarios in associated but not yet revised training materials. XXX MORE?

The issue associated with the latter situation—known sequences with function inferred from the computational analysis of annotated databases—has not been addressed specifically by the CAFC. However, the USPTO announced in 1997 that it would allow claims on cDNA fragments or expressed sequence tags (ESTs) based on their utility as probes.[6] In January 2001, after responding to considerable public debate about the matter, the USPTO published guidelines requiring a specific, substantial, and credible real world utility for every claimed invention. Associated but interim training materials provide examples of contemporary scenarios, including the use of computational analyses of annotated sequences to establish the utility of a claimed EST or cDNA fragment—so-called "genomic patents". However, the guidelines emphasize that utility is evaluated on a case-by-case basis, according to scientific principles, and many remain skeptical of the validity of genomics patents.[7]

---

[4] Antonio Regalado, *The Great Gene Grab*, 103 THE TECHNOLOGY REVIEW 48 (2000) (quoting Professor Rebecca Eisenberg).

[5] David R. Bentley, *Genomic sequence information should be released immediately and freely in the public domain*, 274 SCIENCE 5287 (1996).

[6] John Murray, *Owning Genes: Disputes Involving Dna Sequence Patents*, 75 CHI.-KENT. L. REV. 231, 239 (1999).

[7] Arti K. Rai, *The Information Revolution Reaches Pharmaceuticals: Balancing Innovation Incentives, Cost, and Access In The Post-Genomics Era*, 2001 U. ILL. L. REV. 173, 194 fn100 (2001).

Computational methods are undoubtedly an essential and accepted tool in molecular biology. The patent office, moreover, has been evaluating patent applications that rely on computational methods to describe the claimed sequence and define its utility since at least 1998, and probably for as long as scientists have been using them. Many of these patents have now issued. Nonetheless, whether and how computational methods may be used to establish the patentability of a genetic sequence has not been addressed by the courts, and is not apparent in the legal or scientific literature.

In this paper, I review the law, politics, and administrative procedure relating to "genomic patents"; i.e. patents claiming gene sequences whose utility is based upon similarity to sequences of known function. I then review recently issued patents to assess whether or how computational methods are currently used to (1) describe the claimed gene or nucleotide sequence, and (2) establish the utility of an EST or cDNA fragment. I critique these current practices and respond to criticisms. I find that, in general, the patents are legally and scientifically sound; they may, however, be undesirable for social and political reasons.

II. The Legal and Political Context

To appreciate the use of computational methods in describing and defining the utility of EST patents, some background is necessary. In this section, I provide a simplistic account of the relevant features of patent law and explain why and how genes are patentable. I then consider the written description issue, reviewing the CAFC's assessment of the written description as it applies to gene patents, considering the public's reaction to the ruling that genes are chemical compounds and must be described (preferably by sequence), and summarizing the USPTO's efforts to summarize, update, and implement the law in its *Written Description Guidelines*. To provide context for the debate about patenting ESTs, I next discuss some politics and history. Finally, I consider the utility issue, reexamining a single but important court case, summarizing the USPTO's new *Utility Guidelines*, and noting the public's reaction and predictions about the use of computational methods to define the utility of ESTs.

A. Patent Law and Gene Patents

Patents are issued by the USPTO in accordance with the Patent Statute of 1952 and the courts' interpretations of that statute. An isolated gene sequence is suitable subject matter for a patent, and may be claimed as a "composition of matter." I review the basics of patent law and the logic for patenting gene sequences here.

1. Some Basic Tenets of Patent Law

A patent confers intellectual property rights on an inventor, giving the inventor the right to exclude others from making, using, or selling the claimed

invention for a period of twenty years. Because a patent prevents others from capitalizing on the inventor's ingenuity and investment, it provides the inventor with an incentive to make and develop the invention. However, in order to obtain the temporary monopoly created by a patent, the inventor must disclose the invention. Thus, the patent also assures that new inventions are made available to the public.

The patent application and issued patent comprise a specification and claims. The specification provides background for the invention describes the invention in general and specific terms, and provides examples. It is the technical part of the patent and it tends to be very detailed and comprehensive. The claims are the legal part of the patent. They define the scope of the property claim, much as a surveyor's assessment defines the bounds of a land claim. They are carefully crafted in light of legal precedents and with reference to the invention as disclosed in the specification.

To obtain a patent, the inventor files an application (i.e. a specification and claims) with the United States Patent and Trademark Office (USPTO) and pays certain fees. The application is assessed by an examiner with technical knowledge of the field of the invention. The application must meet criteria established by Congress and clarified by the courts. If the examiner finds that the patent application meets all the applicable requirements, the patent will issue—typically about twenty-four months after the application was filed.

The criteria were established by Congress, acting under the explicit authority of the United States Constitution,[8] in the Patent Act of 1952. Under this statute a patent may be obtained for any (1) process, (2) machine, (3) manufacture, or (4) composition of matter,[9] so long as it satisfies the requirements of (a) utility, (b) novelty, (c) nonobviousness, and (d) description. That is, the invention must have real world utility;[10] it must be novel or new[11] and nonobvious in light of the prior art;[12] and there must be a written description of the invention that shows the inventor's possession of the claimed invention and is sufficient to enable others to practice it.[13]

The patent system is neutral with respect to technology; that is, the same norms apply to all types of inventions. Nonetheless, the USPTO and the CAFC may determine how the general rules will apply to particular areas such as biotechnology and gene patents. It is not uncommon for the USPTO and the CAFC to differ in their interpretations of the statute. At least one scholar has

---

[8] United States Constitution, Art. I, Sect. 8[8]. [The Congress shall have power] To promote the Progess of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries."

[9] 35 U.S.C. §101 (1998). "Whoever invents any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof, may obtain a patent therefor, subject to the conditions and requirements of this title." *Id.*

[10] *Id.*.

[11] 35 U.S.C. §102.

[12] 35 U.S.C. §103.

[13] 35 U.S.C.§112¶1. "The specification shall contain a written description of the invention, and of the manner and process of making and using it, in such full, clear, concise, and exact terms as to enable any person skilled in the art to which it pertains, or with which it is most nearly connected, to make and use he same. . . ."

advocated that the CAFC defer to the informed and technically competent opinion of the USPTO,[14] but the CAFC officially has the final word.[15]

The dynamic typically begins at the administrative level. The USPTO develops policies and sometimes publishes guidelines to be used by patent examiners in assessing patent applications. Both are based on the statute and the CAFC's previous decisions. An inventor may appeal a decision of the examiner to the Board of Patent Appeals and, thereafter, to the CAFC. Disputes arising over patent rights are also taken to a federal district court and, thereafter, to the CAFC. If the CAFC disagrees with the USPTO's decision, the USPTO must revise its policies so that they are in line with the views of the CAFC.

2. The Patentability of Genes

Many people object to the idea of gene patents, arguing that genes are natural and therefore should not be "owned" by anyone.[16] Others object to the consequences of gene patents, arguing that restrictions on access to genetic tools will impede the progress of research.[17] Some seek to limit gene patents to "process" rather than "composition of matter" claims. Irrespective of these public sentiments, policy concerns, and suggestions, isolated genes are simply not per se unpatentable, in any way. However, the information content of genes is probably unpatentable.

In 1980, in the seminal case of *Diamond v. Chakrabarty*, the United States Supreme Court found that genetically engineered bacteria were patentable.[18] The Court cited the Congressional Report accompanying the 1952 Patent Act when it said that the subject matter of patents was meant to include "anything under the sun that is made by man".[19] Thus, the key to the patentable of naturally occurring products of nature is human intervention. Genetic engineering had created organisms whose genomes were manipulated "by man"; therefore, those organisms were patentable.

Eleven years later, in the important case of *Amgen v. Chugai*, the CAFC established that it would treat DNA as a chemical compound: "A gene is a chemical compound, albeit a complex one."[20] Chemicals may be claimed as a

---

[14] Arti K. Rai, *Intellectual Property Rights in Biotechnology: Addressing New Technology*, 34 WAKE FOREST L. REV. 827 (1999).

[15] Rarely, intellectual property cases may be appealed to the United State Supreme Court, whose opinion trumps the opinion of the CAFC. Moreover, the Supreme Court ruled on several important issues in patent law prior to the creation of the CAFC in 1982. In some cases, the Supreme Court and the CAFC have held distinctly different opinions and have ignored the previous decisions of the other court.

[16] *See, e.g.*, Mark Christopher Farrell, *Designer DNA for Humans: Biotech Patent Law Made Interesting for the Average Lawyer*, 35 GONZ. L. REV. 515, 529 (1999/2000) (asserting the common view that "[l]egal protection for the mere discovery of a genetic code sequence already existing in nature seems incorrect."); Murray, *supra* note 6 (providing a general review of gene patenting controversies).

[17] GET CITATION.

[18] *Diamond* v. *Chakrabarty*, 447 U.S. 303 (1980).

[19] *Id.* at 309.

[20] Amgen, Inc. v. Chugai Pharmaceutical Co., 927 F.2d 1200, 1206 (Fed. Cir. 1991).

composition of matter if they are "made by man"—i.e. created in the lab or isolated from nature. In general, matter in its naturally occurring state cannot be patented, but isolated and purified "products of nature" are eligible for patent protection. Thus, it is now clear that "a DNA sequence itself is not patentable. . . [but a] purified DNA molecule isolated from its natural environment . . . is a chemical compound and is patentable if all the statutory requirements are met."[21]

Some people advocate that patent claims involving DNA should be limited to applications or methods of using the DNA; i.e. that patents on the DNA as a composition of matter should not be allowed.[22] However, there is no basis in law for such a limitation on gene patents. As the USPTO recently noted, "Patentable subject matter includes both "process[es]" and "composition[s] of matter." . . . [and p]atent law provides no basis for treating DNA differently from other chemical compounds that are compositions of matter."[23]

For strategic reasons, patents that claim isolated genes as compositions of matter are preferred to patents that claim a particular process for making or using a DNA sequence. A process patent gives the patentee the right to prevent others from using that particular process, but it cannot be used to prevent others from making the resulting product in other ways. However, "a patent on a product per se will be infringed by a competitor making the same product—no matter what process is used to make that product," as was found in the recent case of *Amgen v. Hoescht*.[24] Moreover, a composition patent can be used to prevent others from using the product in any way whatsoever, " . . . even if the inventor disclosed only a single use for the composition."[25]

In short, genes that have been isolated may be patented as a composition of matter, and such patents are extremely powerful weapons in the business world. It is probably not possible, though, to patent pure genetic information.[26] For example, patents on sequences as information stored on a computer readable medium would prevent storage and retrieval of the information. Such patents are unlikely to ever issue, in part because electronic compilations of data are not patentable.[27]

The policy of patents grants the inventor a monopoly in exchange for public disclosure of the invention. Prof. Eisenberg, a noted authority on biotech law, concludes that "[p]atent claims on DNA sequences as "compositions of matter" give patent owners exclusionary rights over tangible DNA molecules and constructs, but do not prevent anyone from perceiving, using, and analyzing information about what the DNA sequence is." Thus, once a patent issues on an

---

[21] U.S. Patent and Trademark Office, Utility Examination Guidelines, 66 Fed. Reg. 1092, 1094 (January 5, 2001), available at http://wais.access.gpo.gov [hereinafter Utility Guidelines].

[22] Utility Guidelines, supra note 21, at 1094-95.

[23] *Id.*.

[24] Amgen, Inc. v. Hoechst, 126 F. Supp. 2d 69 (D. Mass, 2001); *see also* Jennifer Van Brunt, *The Next Move in the Patent Game*, Signals Magazine (April 4, 2001), http://www.signalsmag.com/signalsmag.nsf (discussing the import role of composition of matter gene patents in the business world).

[25] Utility Guidelines, *supra* note 21, at 1095.

[26] Rebecca S. Eisenberg, *Re-Examining the Role of Patents in Appropriating the Value of DNA Sequences*, 49 EMORY L.J. 783 (2000).

[27] *Id.* at 790.

isolated sequence, the information content of that sequence is freely available, "subject only to the inventor's right to exclude others from making, using, and selling the claimed materials."[28]

B. The Written Description Issue: Inferring Structure from Function

Section 112 of the Patent Act sets forth the requirements for the specification, and says that it "shall contain a written description of the invention." This seemingly simple requirement has been interpreted by the courts to require a description that is sufficient to indicate that the inventor had "possession" of the invention.[29] That is, the inventor must fully set forth the claimed invention, providing "sufficient detail that one skilled in the art can clearly conclude that the inventor invented the claimed invention."[30]

An inventor who has reduced his or her invention to practice is clearly in possession of it and will easily satisfy the written description requirement by describing what was done. If an inventor has merely conceived an invention, the inventor must clearly demonstrate conception in order to show possession and satisfy the description requirement. When an invention is not obvious in light what is described, the requirement is not satisfied. Thus, the written description requirement is often intertwined with the issue of obviousness.

I review here the CAFC's early rulings on the description of genes and their obviousness in light of knowledge of an amino acid sequence, and I consider some criticisms of the court's approach and findings. I then summarize the recent *Guidelines* developed by the USPTO and show how they sensibly address most of the concerns.

1. The Court's Interpretation of Gene Descriptions

In the early 1990s, the CAFC determined that a "biomolecule sequence described only by a functional characteristic, without any known or disclosed correlation between that function and the structure of the sequence, normally is not a sufficient identifying characteristic for written description purposes, even when accompanied by a method of obtaining the claimed sequence."[31] That is, a claim to a nucleotide sequence could not be supported by merely naming the protein for which it codes and a method for isolating it.

The court first addressed the issue in 1991 in the case of *Amgen v. Chugai*.[32] It considered the validity of Amgen's patent claim to a "purified and isolated DNA sequence consisting essentially of a DNA sequence encoding human

---

[28] *Id.* at 787.

[29] Vas-Cath Inc. v. Mahurkar, 935 F.2d 1555, 1563-64 (Fed. Cir. 1991) (to satisfy the written description requirement, the specification must "reasonably convey to the artisan that the inventor had possession at that time of the ... claimed subject matter.").

[30] Lockwood v. American Airlines, 107 F.3d 1565,1572 (Fed. Cir. 1997)

[31] U.S. Patent and Trademark Office, Guidelines for Examination of Patent Applications Under the 35 U.S.C. 112, para. 1, "Written Description" Requirement, 66 Fed. Reg. 1099, 1108 fn14 [hereinafter Written Description Guidelines].

[32] Amgen, Inc. v. Chugai Pharmaceutical Co., 927 F.2d 1200 (Fed. Cir. 1991).

erythropoietin."[33] Amgen had not isolated and sequenced the gene and the polypeptide sequence of human erythropoietin was unknown. The court decided that knowing a method to isolate and sequence the gene was not enough—Amgen needed to know and describe the sequence; that is, it needed to actually reduce the invention to practice.[34]

The court based this decision on its assessment of the DNA as a chemical compound. It noted that "conception of a chemical compound requires that the inventor be able to define it so as to distinguish it from other materials."[35] It then concluded that "[i]t is not sufficient to define [the erythropoietin gene] solely by its principal biological property, e.g., encoding human erythropoietin, because an alleged conception having no more specificity than that is simply a wish to know the identity of any material with that biological property."[36] Rather, the inventor must have "a mental picture of the structure of the chemical, or [be] able to define it by its method of preparation, its physical or chemical properties, or whatever characteristics sufficiently distinguish it.[37]

The court addressed the issue again in 1993 in the case of *Fiers v. Revel*.[38] In this interference action between parties seeking similar but as yet unissued patents, the court addressed the validity of a potential claim to a "DNA which consists essentially of a DNA which codes for a human fibroblast interferon-beta polypeptide.[39] The court cited *Amgen* in holding that "conception of any chemical substance, requires a definition of that substance other than by its functional utility" and then elaborated that "[c]onception of a substance claimed per se without reference to a process requires conception of its *structure, name, formula, or definitive chemical or physical properties*" (emphasis added).[40] In short, the court found that "[a]n adequate written description of a DNA requires . . . a description of the DNA itself."[41]

At about the same time that it was addressing the written description requirement as applied to gene patents, the CAFC addressed the issue of the obviousness of a DNA sequence when the amino acid sequence of the polypeptide for which it codes is already known.[42] To the surprise of many biologists, the CAFC determined that knowing the amino acid sequence of a polypeptide and a

---

[33] *Id.* at 1204.
[34] *Id.* at 1206.
[35] *Id.*
[36] *Id.*
[37] *Id.*
[38] Fiers v. Revel, 984 F.2d 1164 (Fed. Cir. 1993).
[39] *Id.* at 1166.
[40] *Id.* at 1169. The court elaborate on the connection between conception and description, noting that "[i]f a conception of a DNA requires a precise definition, such as by structure, formula, chemical name, or physical properties, as we have held, then a description also requires that degree of specificity. To paraphrase the Board, one cannot describe what one has not conceived." *Id.* at 1171.
[41] *Id.* at 1171.
[42] An invention must be nonobvious to qualify for a patent. *See* note 12, *supra*, and accompanying text; Jeffrey S. Dillen, *DNA Patentability - Anything but Obvious*, 1997 WIS. L. REV. 1023 (1997) (reviewing case law related to the issue of the obviousness of a DNA sequence if the amino acid sequence for which it codes is known).

general method of cloning does not make the naturally occurring nucleotide sequence obvious. The logic is, however, consistent with the court's assessment of the written description requirement as it applies to claims to DNA.

The court first addressed the issue of obviousness in 1993 in a case called *In re Bell.*[43] Bell sought to claim the sequences which code for *human* insulin-like growth factors (IGF) I and II; the amino acid sequence of these proteins was already known. The CAFC again focused on the DNA molecules as chemical compounds rather than assessing the methods used to isolate the DNA. The court acknowledged that, "knowing the structure of the protein, one can use the genetic code to hypothesize possible structures for the corresponding gene" but it also acknowledged the vast number of sequences that could code for a protein.[44] Because it was not known which of the possible sequences would be found *in humans*, the court found that the human sequence was not obvious.[45]

The court addressed the issue again in 1995 in a case called *In re Deuel*[46] Deuel claimed: "A purified and isolated DNA sequence consisting of a sequence encoding human heparin binding growth factor of 168 amino acids having the following amino acid sequence: Met Gln Ala ... [remainder of 168 amino acid sequence]."[47] The court saw that the claim was "tantamount to the general idea of all genes encoding the protein, all solutions to the problem." And it wisely acknowledged that this set of sequences "might have been obvious from the complete amino acid sequence of the protein, coupled with knowledge of the genetic code" explaining that "this information may have enabled a person of ordinary skill in the art to envision the idea of, and, perhaps with the aid of a computer, even identify all members of the claimed genus."[48] However, because the amino acid sequence was previously unknown, the court found that the claim was not invalid for obviousness.

These rulings of the CAFC may be summarized as follows: A claim to a DNA must describe the DNA; it cannot be inferred by naming the protein for which it codes and a method for isolating the DNA. Even if the amino acid sequence of the protein is known, the actual sequence that codes for the protein in a particular organism is not. Therefore, the DNA sequence must be established to claim a gene specific to a particular organism. However, if the amino acid sequence of the protein is newly discovered, then the entire class of DNAs that could code for the protein is also newly discovered. In this case, a set or "genus" of DNA sequences may be claimed by acknowledging the genetic code and describing the polypeptide sequence.

In 1997, the court expanded these precedents to address the description of a set, or genus, of DNAs (rather than a single molecule, or species) in a case known commonly at *U.C. v. Eli Lilly.*[49] The University of California sought to claim

---

[43] In re Bell, 991 F.2d 781 (Fed. Cir. 1993).

[44] *Id.* at 784 (Fed. Cir. 1993). The court acknowledged the possibility that a known amino acid sequence is specified exclusively by unique codons, in which case the gene would be obvious. *Id.*

[45] *Id.*

[46] In re Deuel, 51 F.3d 1552 (Fed. Cir. 1995).

[47] *Id.* at 1555.

[48] *Id.* at 1560.

[49] Regents of University of California v. Eli Lilly, 119 F.3d 1559 (Fed. Cir. 1997).

mammalian and vertebrate insulin cDNA based upon a description of human insulin cDNA. The court found the description inadequate. It said that a "written description of an invention involving a chemical genus, like a description of a chemical species, "requires a precise definition, such as by structure, formula, [or] chemical name," of the claimed subject matter sufficient to distinguish it from other materials."[50] It concluded that "a generic statement such as "vertebrate insulin cDNA" or "mammalian insulin cDNA," without more, is not an adequate written description of the genus because . . . it does not define any structural features commonly possessed by members of the genus that distinguish them from others."[51]

The *Lilly* court asserted that DNA claims would require "a kind of specificity usually achieved by means of the recitation of the sequence of nucleotides that make up the DNA" and, by analogy, that claims to a genus of cDNAs would require reciting a "representative number of cDNAs, defined by nucleotide sequence." The court refused, however, to "speculate in what other ways a broad genus of genetic material may be properly described ...."[52]

2. The Public Debate about Treating DNA as a Chemical Compound

The courts' treatment of genes as chemical compositions has been debated extensively, both as it relates to the issue of obviousness and the written description.[53] By treating DNA as a chemical, the CAFC has simultaneously lowered the bar for non-obviousness (by finding that knowledge of an amino acid sequence and a general method for identifying genes with the use of nucleotide probes does not make the DNA sequence obvious) and raised the bar for the written description (by requiring that genes are actually isolated and sequenced before being patented).[54]

Rai, for example, argues that the CAFC's treatment of DNA as a subset of chemical technology is "fundamentally misconceived" and reflects the court's failure to recognize DNA-based technologies "as involving information first and foremost."[55] She says that, as a result, "the courts have thereby made patent protection too strong in some respects and too weak in others."[56] Eisenberg also emphasizes the importance and value of DNA sequences as information.[57] She finds that "the chemical analogy is of little value as a strategic guide to exploiting this information as intellectual property."[58]

---

[50] *Id* at 1568.
[51] *Id.*
[52] *Id.* at 1569. START HERE
[53] *See, e.g.*, Todd R. Miller, *Motivation and Set-Size: In Re Bell Provides a Link Between Chemical and Biochemical Patent Claims*, 2 U. BALT. INTELL. PROP. J. 89 (1993) (drawing upon and citing previous participants in the debate).
[54] *See* Part II.B.1; *see also* Rai, *supra* note 14.
[55] Rai, *supra* note 14, at 836 ("Although DNA is, obviously enough, a chemical compound, it is more fundamentally a carrier of information.").
[56] *Id.*
[57] Eisenberg, *supra* note 26.
[58] *Id.* at 785.

There are defenders of the court's approach. Margaret Sampson suggests that the heightened description approach helps prevent overly broad patents.[59] Sampson argues that the heightened description requirement prevents an inventor from restricting the use of "homologs, alleles, polymorphisms, and isoforms found in the same gene family, all of which have a high degree of sequence identity with the gene, but not 100% identity," and limits the ability of inventors to assert rights to sequences of which they have no knowledge, in organisms with which they have never worked.[60] As discussed in Part III.B.2, this does not appear to be the case.

Perhaps more importantly, the court's approach may be good policy if it encourages inventors to establish nucleotide sequences for known proteins and prevents them from asserting rights to genes without ever revealing their sequences. Indeed, by treating DNA as a chemical compound and requiring inventors to describe its structural attributes, the court has effectively required inventors to (1) determine the critical information attribute of a DNA (i.e. the nucleotide sequence) and (2) reveal it to the public. These rulings may therefore promote the discovery of genetic information—by providing an incentive to discover gene sequences, as well as the dissemination of genetic information—by requiring that the information is revealed to the public in the patent.

3. The PTO's Guidelines for Examination of the Written Description

The USPTO published its *Guidelines for Examination of Patent Applications Under the 35 U.S.C. 112, para. 1, "Written Description" Requirement* (Written Description Guidelines) on January 5, 2001. This document reflects the USPTO's understanding of the law on the statutory requirement of a written description, and was created to provide guidance to the examiners who must evaluate patent applications in light of the law. An interim version of the document was previously made available to the public for comments; in the final version, the USPTO summarizes and responds to those comments, but does not change the guidelines substantially. The document provides a comprehensive, accurate, and accessible summary of the law, and indicates how the USPTO has applied and will apply the law—at least until the CAFC contradicts its interpretation.

The Written Description Guidelines provides a sensible restatement of the law, noting that "[a]n adequate written description of the invention may be shown by any description of sufficient, relevant, identifying characteristics so long as a person skilled in the art would recognize that the inventor had possession of the claimed invention."[61] It also acknowledged the finding of the *Amgen* court, i.e. when "an invention is described solely in terms of a method of its making coupled with its function and there is no described or art-recognized correlation or

---

[59] Margaret Sampson, *The Evolution of the Enablement and Written Description Requirements Under 35 U.S.C. 112 in the Area of Biotechnology*, 15 BERKELEY TECH. L.J. 1233, 1261 (2000).
[60] *Id.*
[61] Written Description Guidelines, *supra* note 31, at 1105.

relationship between the structure of the invention and its function," the description is inadequate.[62]

According to the Written Description Guidelines, an invention may be sufficiently described by disclosure of "complete or partial structure, other physical and/or chemical properties, *functional characteristics when coupled with a known or disclosed correlation between function and structure*, or some combination of such characteristics (emphasis added).[63] For at least some biomolecules, such characteristics include "a sequence, structure, binding affinity, binding specificity, molecular weight, and length" but "other identifying characteristics or combinations of characteristics may demonstrate the requisite possession".[64]

Thus, the Written Description Guidelines acknowledge that molecules may be described not only by sequence, but also by functional attributes when such attributes are clearly associated with structural attributes. Indeed, the *Guidelines* instruct examiners to consider "the level of skill and knowledge in the art, partial structure, physical and/or chemical properties, [as well as] *functional characteristics alone or coupled with a known or disclosed correlation between structure and function.*"[65]

The Written Description Guidelines also address the court's interpretation of the written description as it applies to a claimed genus, noting that a claim to a genus is satisfied "through sufficient description of a representative number of species by actual reduction to practice. . ., reduction to drawings. . . , or by disclosure of relevant, identifying characteristics." The Written Description Guidelines again indicate that such characteristics include "structure or other physical and/or chemical properties, . . . *functional characteristics coupled with a known or disclosed correlation between function and structure*, . . .[and a combination of such identifying characteristics, ..."

The most novel and interesting direction in the Written Description Guidelines pertains to the adequacy of the description of a genus of DNAs by reference to an amino acid sequence. The USPTO notes two comments asserting that, "if the amino acid sequence for a polypeptide whose utility has been identified is described, then the question of possession of a class of nucleotides encoding that polypeptide can be addressed as a relatively routine matter using the understanding of the genetic code." The suggestion was incorporated into the Written Description Guidelines as follows: "if an applicant disclose[s] an amino acid sequence, it [is] unnecessary to provide an explicit disclosure of nucleic acid sequences that encode[] the amino acid sequence. Since the genetic code is widely known, a disclosure of an amino acid sequence . . . provide[s] sufficient information such that one would accept that an applicant was in possession of the full genus of nucleic acids encoding a given amino acid sequence, but not necessarily any particular species."[66]

---

[62] *Id.*.
[63] Written Description Guidelines, *supra* note 31, at 1106.
[64] Written Description Guidelines, *supra* note 31, at 1110 fn 42.
[65] Written Description Guidelines, *supra* note 31, at 1106.
[66] Written Description Guidelines, *supra* note 31, at 1111 fn 57.

The Written Description Guidelines note, though, that "this does not mean that applicant was in possession of any particular species of the broad genus."[67] Such claims may therefore be allowed, but may fail to preclude subsequent claims to sequences that are, e.g., specific to a particular organism.

C. The Politics of EST Patents

In the early 1990s, when the courts were assessing the legal implications of claiming DNA whose sequence was not yet known, scientists were beginning to produce large numbers of cDNA fragments known as expressed sequence tags, or ESTs. These short nucleic acid sequences were relatively easily discovered, but their function was usually unknown—in sharp contrast to the situation of the previous decade, when sequences of known function were sought and obtained after substantial focused effort.

The community was divided about the merits of patenting ESTs.[68] The National Institute of Health and then Craig Ventor sought to patent them, but the Human Genome Organization (HUGO) vehemently opposed any and all such efforts.[69] XXXMORE HUGO believed that ESTs were research tools, and thought they and all sequences should be viewed as part of pre-competitive information.[70] Nonetheless, by 1996, the USPTO was deluged with over half a million applications for patents on ESTs. At that point, the office stopped tracking them.[71]

Fortunately for the USPTO, the flood abated, with the number of EST patent applications dropping dramatically around 1998.[72] Various PTO officials have characterized three cycles or generations of EST patents: The first generation comprises applications that do not disclose the gene associated with the EST. The second generation comprises applications where the function of the protein being expressed by the gene is determined by homology searches. In the third-generation patents, "[the inventors] have actually found the function by doing the science," piecing together the complete open reading frame (ORF) for the gene. In April 2001, it was estimated that the PTO had received as many as 25,000 third generation applications.[73] XXXCHECK

The arguments about patenting ESTs have focused on utility. As Professor Eisenberg noted in 1992, "the argument against allowing NIH to patent the sequences is not really that these sequences are useless, but rather that NIH does not yet know what they are good for and should not be able to claim patent rights ahead of subsequent researchers who figure it out. It is the as yet undiscovered

---

[67] Written Description Guidelines, *supra* note 31, at 1102.

[68] Gary Zweiger provides a cogent and timely review of the history of genomics, including an assessment of the companies and individuals who sought to patent ESTs and those who opposed such business tactics. Gary Zweiger, TRANSDUCING THE GENOME: INFORMATION, ANARCHY, AND REVOLUTION IN THE BIOMEDICAL SCIENCES (2001). *See also* Murray, *supra* note 6.

[69] Human Genome Organization (HUGO), *Statement on Patenting of DNA sequences - In Particular Response to the European Biotechnology Directive* (April 2000).

[70] *Id.*

[71] Van Brunt, *supra* note 24.

[72] *Id.*

[73] Todd Dickinson. Comments at Congressional Hearing on Genomic Inventions, *supra* note 1; Van Brunt, *supra* note 24.

utility of the sequences, rather than the uses that are disclosed in the patent application, that makes NIH's patent claims worth fighting about.[74] The general thinking is that ESTs should be patentable if the full gene sequence and its function are known.[75] If so, the first generation EST patent applications will not satisfy the utility requirement, but the third generation applications will.

Patent applications for ESTs in the second generation, where utility is inferred from the computational analysis of genomic databases, are most difficult to assess. The Director of the USPTO explained to members of Congress in July 2000: "The question comes down to ... how much utility can be inferred from the computer modeling that is used now to determine the utility associated with a particular EST. The question is what percentage of that analogous information—it's called percent homology in the term of the art—is sufficient, in order to justify the utility."[76] In short, the question is whether a finding of homology of an EST with a known gene is sufficient to establish utility, and hence patentability, of the EST.

The second generation EST patents are politically contentious because they provide patent rights to early stage research tools. Such patents could affect both the pace of genetics research and the structure of industry. If the patenting of ESTs restricts researchers' access to them, such patents could impede complete characterization of genes and delay or restrict exploration of genetic materials for the public good.[77] Whether or not this is true may depend upon the business methods adopted in the relevant industries. For example, the use of non-exclusive licenses and the creation of patent pools could facilitate the widespread use of patented ESTs.[78] On the other hand, such patents may provide incentives for research and development of gene fragments, and could foster the development of companies that specialize in genomics research.

Randall Scott of Incyte, a company that focuses on the accumulation and analysis of early stage research information, argues for EST patents—even when the precise biological activity of the gene is unknown. Scott rightly emphasizes that "a patent should be rewarded for commercial utility, not for biological function, and there's an important distinction."[79] He argues that ESTs are useful "as tools, as diagnostics, as markers for disease and drug therapy," and such uses do not require knowledge of their biological function. Thus, he says, "the real world utility of genes is not just buried in their biological function and what they do naturally in the body."[80]

---

[74] Rebecca S. Eisenberg, *Genes, Patents, and Product Development*, 257 SCIENCE 5072 (1992).
[75] *See, e.g.*, Murray, *supra* note 6 (1999).
[76] Todd Dickinson. Comments at Congressional Hearing on Genomic Inventions, *supra* note 1.
[77] Murray, *supra* note 6 at 254.
[78] *But see* Rai, *supra* note 17 (critiquing the idea that the market can compensate for the blocking effect of patents on early stage research tools).
[79] Dr. Randal W. Scott, President And Chief Scientific Officer, Incyte Genomics. Statement at Congressional Hearing on Genomic Inventions, *supra* note 1 (noting as an example that the common indicator of prostate cancer is the observation of a certain protein in the blood; the function of the protein is unknown, but tests for the protein clearly have significant commercial utility).
[80] *Id.*

Scott's view contrasts sharply with the view of officials at Genentech, a company that is involved in the development of pharmaceutical products. Genetech officials believe that "the utility of a particular gene or protein cannot be known unless one has determined its [biological] function."[81] And such determination requires laboratory research, not genomic analysis. Dennis Henner, of Genentech, told members of Congress that "computer modeling is not sufficiently accurate to predict protein function based solely on gene comparisons."[82] Therefore, he said, "the utility of a particular gene or polypeptide rarely can be demonstrated until there has been a sufficient characterization of the function of a gene or its expression product . . . through relevant biological assays."[83]

D. The Utility Issue: Inferring Function from Structure

Section 101 of the Patent Act establishes that "[w]hoever invents any . . . *useful* . . . composition of matter . . . may obtain a patent therefor . . .".[84] This so-called utility requirement historically was and, in many cases still is, trivial. In 1817 it was interpreted to mean only that an invention could not be mischievous or immoral. Today, the utility requirement reflects more general policy concerns. Utility became an issue in the chemical arts in 1966, when the court ruled that a chemical compound with no known practical use could not be patented. It is now a major issue in the patenting of ESTs.

The issue of the utility of ESTs implicates the validity of structure-function relationships in biochemistry, and the consequences of such patents for further discoveries relating to the associated gene. As the Director of the USPTO acknowledged in July 2000, "legitimate questions have been raised about just what genomic discoveries, if any, should be patentable and whether genomic patents will inhibit researchers' access to the data, materials, and methods needed to develop new tools for the diagnosis and treatment of disease"[85]

In the section, I review the courts' general rulings on utility, and the USPTO's guidelines for applying the utility requirement to biotech inventions. I consider

---

[81] Dennis J. Henner, Ph.D., Senior Vice President, Research, Genentech, Inc. Statement at Congressional Hearing on Genomic Inventions, *supra* note 1.

[82] *Id.*

[83] *Id.* He elaborated as follows:

> The degree of homology can be an important indicator that the sequence being analyzed is similar to, or within a class of known proteins based on the degree of identity it shares with the known sequence. . . . Homology analysis, however, is a limited tool for predicting results. In our experience, homology analysis, standing alone, is not a sufficiently reliable indicator to base scientific or business decisions upon. . . . Accordingly, where a particular biological activity is the only basis for the utility of a particular gene or expression product, a homology-based prediction should not be capable of satisfying the requirements of our law in a majority of situations.

Id.

[84] 35 U.S.C. §101 (1998); *see also* note 9, *infra*, and accompanying text.

[85] Todd Dickinson, Under Secretary of Commerce for Intellectual Property and Director of the United States Patent and Trademark Office. Prepared Statement at Congressional Hearing on Genomic Inventions, *supra* note 1.

public commentary and attempt to determine (1) whether or when, under the guidelines, an inventor must know the biological function of the protein coded by the gene associate with the claimed EST, and (2) whether the inventor can establish that function by analyzing sequence similarity to genes of known function.

1. The Courts' Interpretation of Utility in the Chemical Arts

No court has yet addressed the application of the utility requirement to partial nucleotide sequences. Thus, it is possible that the opinions of academics and the policies of the USPTO will be found irrelevant and inapplicable, respectively. The USPTO has purportedly arranged for two interested parties to present the issue of the utility of ESTs to the court in a "test case."[86] In July 2000, this case was purportedly set to go to the Board of Appeals; if so, it could appear before the CAFC as early as January 2002.[87]

The United State Supreme Court did, however, address the issue of utility as it applies to the chemical arts in the 1966 case of *Brenner v. Manson*.[88] Manson had devised a method for making a certain steroid compound. The Court found that Manson had failed to assert any utility for the process, other than its use in research by chemists. Because the invention did not have practical benefits for the public, and because a patent on the process could "confer power to block off whole areas of scientific development, without compensating benefit to the public," the court found that it failed to meet the utility requirement.[89] In summary, the Court declared that "a patent is not a hunting license. It is not a reward for the search, but compensation for its successful conclusion."[90]

The *Brenner* court explicitly rejected Manson's argument for utility based upon the observation that a compound similar to the one produced by his process (an "adjacent homologue") had been shown to inhibit the growth of tumors in mice. The USPTO had found that Manson had not disclosed "a sufficient likelihood that the steroid yielded by his process would have similar tumor-inhibiting characteristics," and the Court accepted its finding.[91] In short, because Manson had failed to provide a convincing argument for the function of the steroid based upon its structural similarities to compounds with known functions, he had failed to assert a practical utility.

The Court's reliance on the USPTO's determination that Manson could not reasonably infer the function of his steroid from its structure is important. It suggests that assertions for the utility of ESTs based upon their structural similarity to genes coding for proteins of known function depends upon the USPTO's determination of the scientific validity of such an inference.

2. The PTO's Guidelines for Examination of Utility

---

[86] *Id.*
[87] *Id.*
[88] 383 U.S. 519 (1966).
[89] *Id.* at 534.
[90] *Id.* at 536.
[91] *Id.* at 532.

The USPTO published its *Utility Examination Guidelines* (Utility Guidelines)[92] on January 5, 2001. This document reflects the USPTO's understanding of the law on the statutory requirement of utility and was created to provide guidance to the examiners who must apply it.[93] As for the Written Description Guidelines, the USPTO summarizes and responds to comments on a previously published version.[94]

The Utility Guidelines have been more contentious than the Written Description Guidelines, because the utility of ESTs is the key factor in assessing their patentability.[95] Excellent synopses of the document, with critical commentary, are already available.[96]

The Utility Guidelines require the inventor to identify a *specific*, *substantial*, and *credible* utility for the claimed invention, unless such a utility is already well established.[97] This three-part test raises the bar for showing utility because previous guidelines required only a credible utility. However, a "specific" and "substantial" utility has been required by the courts. Thus, the new guidelines are more in line with case law than previous guidelines.

An asserted utility is *credible* unless (1) the logic underlying the assertion is seriously flawed, or (2) the facts upon which the assertion is based are inconsistent with the logic underlying the assertion. The credibility of an asserted utility is assessed from the standpoint of a person of ordinary skill in the art, but the presumption favors the inventor. For example, since at least some nucleic acids can be used as probes, chromosome markers, or diagnostic markers, the assertion that any particular DNA can be used in this is accepted.

An asserted utility is *substantial* if it defines a "real world" use. If further research is required to confirm or identify the use, the use is not substantial. Thus, claims that a nucleic acid is useful for studying the properties of the gene itself are not substantial.

---

[92] Utility Guidelines, *supra* note 21.

[93] The USPTO emphasizes this point in the Utility Guidelines, clarifying that it is not free to develop its own rules about the patentability of DNA. Utility Guidelines, *supra* note 21, at 1095 ("The USPTO must administer the laws as Congress has enacted them and as the Federal courts have interpreted them. Current law provides that when the statutory patentability requirements are met, there is no basis to deny patent applications claiming DNA compositions, or to limit a patent's scope in order to allow free access to the use of the invention during the patent term.").

[94] U.S. Patent and Trademark Office, Revised Interim Utility Guidelines, 64 Fed. Reg. 71440 (Dec. 21, 1999); correction at 65 Fed. Reg. 3425 (Jan. 21, 2000).

[95] Expressed Sequence Tags are "patentable to the same extent that any other invention is patentable, so long as they meet the test of patentability. And the question that it basically comes down to . . . is the question of utility and the ability to demonstrate sufficient utility to meet the section 101 standard." Todd Dickinson. Comments at Congressional Hearing on Genomic Inventions, *supra* note 1.

[96] *E.g.* Timothy A. Worrall, *The 2001 PTO Utility Examination Guidelines and DNA Patents*, 16 BERKELEY TECH. L.J. 123 (2001); *The Fate of Gene Patents Under the New Utility Guidelines*, 2001 Duke L. & Tech. Rev. 0008 (2001).

[97] Utility Guidelines, *supra* note 21; *see also* U.S. Patent and Trademark Office, Revised Interim Utility Guidelines Training Materials, *available at* http://wais.access.gpo.gov [hereinafter Utility Training Materials]; Todd Dickinson. Comments at Congressional Hearing on Genomic Inventions, *supra* note 1.

An asserted utility if *specific* when it is particular to the subject matter claimed. For example, asserting that an EST is useful as a "gene probe" or "chromosome marker" in not sufficiently specific; the inventor must disclose a particular gene for the probe, or chromosome target for the marker. By the same logic, asserting that an EST has diagnostic utility is typically insufficient; the inventor must identify the condition that is diagnosed.

The Utility Guidelines are widely viewed as having raised the bar on utility as it applies to the patenting of ESTs. However, they appear to clearly indicate that ESTs are patentable, even if the function of the encoded gene product is unknown. They state unequivocally that "[t]he utility of a claimed DNA does not necessarily depend on the function of the encoded gene product. A claimed DNA may have a specific and substantial utility because, e.g., it hybridizes near a disease-associated gene or it has a gene-regulating activity."[98] And they clearly suggest that computational methods such as sequence comparisons may be used to identify the relevant gene and thereby provide the required specific utility.

3. The Public Debate about Using Genomics to Establish Utility

In July 2000, Todd Dickinson told members of Congress that officials at the USPTO believed the new "heightened standard of utility w[ould] allow appropriate patents on genomic inventions, while also resulting in the rejection of hundreds of genomic patent applications, particularly those that only disclose *theoretical utilities*"[99] (emphasis added). As one reporter described it, researchers "take a gene, or even just a piece of a gene, plug it into a computer, and instantly turn up vast amounts of intriguing but *theoretical* information about it"; they then file for patents "without doing a single experiment or 'getting [a] pipette wet'".[100]

These comments reflect a not uncommon sentiment that knowledge acquired by experimentation in the lab is superior to knowledge acquired through the analysis of databases. John Golden recently argued that "the science of "bio-informatics" [is] still in its infancy, [and] current computer-based methods for studying genetic sequences have failure rates as high as 95%."[101] He objected to the USPTO's idea that "computer-based analogy to a known useful sequence is presumptively sufficient for patentability" and concluded that "installing a presumption in favor of the reliability of computer-based studies could . . . ultimately give away most of what a meaningful utility requirement is meant to protect."[102]

Clearly, assessing the results of database analyses can be difficult and the need to interpret findings that are typically associated with probabilities may be unfamiliar and non-intuitive to scientists who are accustomed to interpreting the typically binary feedback of laboratory results. Nonetheless, even some officials recognize that searching sequence databases for similar genes is common practice

---

[98] Utility Guidelines, *supra* note 21, at 1095.
[99] Todd Dickinson. Statement at Congressional Hearing on Genomic Inventions, *supra* note 1.
[100] Merril Goozner, *Patenting Life*, THE AMERICAN PROSPECT (December 18, 2000).
[101] John M. Golden, *Biotechnology, Technology Policy, and Patentability: Natural Products and Invention in the American System*, 50 EMORY L.J. 101, 188 (2001).
[102] *Id.*

and is "very well established and very well accepted in the academic community."[103]

Patent experts believe the USPTO new Utility Guidelines are unlikely to be overturned by the court, perhaps because the court has traditionally failed to enforce the utility requirement very strictly.[104] Perhaps the more interesting question is whether utilities asserted by database analyses can be justified scientifically.

III. The Use of Computational Techniques in Gene Patents

There is currently no published study of patent office decisions examining claims to ESTs or the implementation of the new Written Description and Utility Guidelines.[105] Thus, it is not known how or to what extent the Guidelines have affected the type or style of patents issuing on ESTs.

In this section, I rely on various searches of recently issued patents, a close reading of more than twenty patents issuing on ESTs, and the examples provided in the USPTO's Training Materials to determine how scientists and their patent attorneys are using computational methods to satisfy the written description and utility requirements. I then critique these uses from both scientific and legal perspectives.

A. Finding EST Patents

It is not known how many patents have issued on genes in general or ESTs in particular, but from all accounts and consistent with all estimates, there are likely tens of thousands of gene patents and hundreds if not thousands of EST patents. Furthermore, there is no easy way to identify a patent as an EST patent, short of reading and considering it in its entirety. I describe here my search approach and some suggestive data on trends in the issuance of patents using computational methods.

1. Methods for Searching the Databases

I used various combinations of key word searches of the Lexis[106] patent database, with various field and date restrictions, to identify a manageable number of recent patents on ESTs that I could examine closely. My search methods were exploratory, and the sample of patents that I chose to examine closely may not be representative of EST patents in general.

---

[103] Martin Enserink, *Patent Office May Raise The Bar on Gene Claims*, 287 SCIENCE 5456 (2000) (citing Doll). The reported statement of an Incyte representative that "Everybody uses these techniques and they are virtually 100% correct" overstates the case and fails to acknowledge the importance of interpreting probabilities. *Id.*

[104] *Id.*

[105] Goozner, *supra* note 100.

[106] This database is available by subscription only; however, all patents examined here are available in their entirety at the USPTO website, http://www.uspto.gov.

There is a classification system for patents, and all patents list one and usually several class/subclass categories. I examined the classification of several patents that I had determined by various means to be EST patents, and found that most (although not all) of them listed Class 536, Subclass 23.1 Class 536 is "organic compounds" and within it, subclass 23.1 is "DNA or RNA fragments or modified forms thereof", and subclasses 23.2 to 23.7 are DNA or RNA fragments that encode a particular type of protein. Thereafter, I restricted my searches to this Class and Subclass.

Given the large number of EST patents, I focused on patents issued most recently in the summer of 2001, between June 1 and August 15.

After examining a number of gene patents, I found that claims to sequences as compositions of matter invariable referred to a sequence given in the specification as "SEQ ID" followed by a identification number. These claims usually also specified that the claimed compound was a "polynucleotide." I therefore restricted my searches to patents with these terms in the claims. I also restricted several searches to claims that included the term "percent identity" for reasons that will be obvious later. The most useful combination of keywords for identifying EST patents, given the previously noted restrictions, were the terms "EST" or "cDNA" in conjunction with "fragment" or "partial".

With these search criteria, I obtained 17 patents, each assigned to one of four companies. Because 11 of the 17 patents were assigned to DuPont and only 2 were assigned to Incyte,[107] I searched for patents issued to various companies prior to June 1, 2001. I include 4 additional patents issued to Incyte, because they have been a vocal participant in the debate about genome patents.[108] I also

---

[107] The patents assigned to Dupont included: Patent Number 6,255,090: Plant aminoacyl-tRNA synthetase (July 3, 2001); Patent Number 6,271,441: Plant aminoacyl-tRNA synthetase (August 7, 2001); Patent Number 6,255,114: Starch biosynthetic enzymes (July 3, 2001); Patent Number 6,252,137: Soybean homolog of seed-specific transcription activator from Phaseolus vulgaris (June 26, 2001); Patent Number 6,242,256: Ornithine biosynthesis enzymes (June 5, 2001); Patent Number 6,262,345: Plant protein kinases (July17, 2001); Patent Number 6,274,379: Plant sorbitol biosynthetic enzymes (August 14, 2001); Patent Number 6,248,584: Transcription coactivators (June 19, 2001); Patent Number 6,251,668: Transcription coactivators (June 26, 2001); Patent Number 20010005749: Aromatic amino acid catabolism enzymes (June 28, 2001); Patent Number 20010010909: Chromatin Associated Proteins (August 2, 2001).

The patents assigned to Incyte were: Patent Number 6,277,568: Nucleic acids encoding human ubiquitin-conjugating enzyme homologs (August 21, 2001); Patent Number 20010010913: Extracellular adhesive proteins (August 2, 2001).

The remaining patents included one patent assigned to Dendreon Corporation: Patent Number 6,194,152: Prostate tumor polynucleotide compositions and methods of detection thereof (February 27, 2001); two assigned to Bayer Corporation: Patent Number 6,262,333: Human genes and gene expression products (July 17, 2001); Patent Number 6,262,334: Human genes and expression products: II (July 17, 2001); and one assigned to a foreign corporation, Zeneca: Patent Number 6,265,560: Human Ste20-like stress activated serine/threonine kinase (July 24, 2001).

[108] The patents are: Patent Number 5,912,130: Human Homolog of the rat G protein gamma-5 subunit (Jun. 15, 1999);Patent Number 5,783,418: Human homolog of the rat G protein gamma-5 subunit (Jul. 21, 1998); Patent Number 5,932,442: Human regulatory molecules (Aug. 3, 1999); Patent Number 5,840,544: DNA encoding rantes homolog from prostate (Nov. 24, 1998).

examined the patent that Incyte claims to be the first issued EST patent,[109] and a patent thought to be an EST patent that issued earlier.[110]

2. Quantitative Search Results

I conducted some systematic searches of patents issued over the past five years to assess temporal trends in the number of EST patents and the use of various computational methods in those patents.

I looked at the temporal variability in patents in patents listing Class 536, Subclass 23.1 and 23.2-.7 (Table 1) to assess trends in the number of EST patents over the last five years. Cursory inspection of patents in subclass 23.1 showed that not all but many of the patents listing this subclass were EST patents. The number of patents in these classes increased about three-fold from 1996 to 1998, and then remained fairly constant, with an average of 175 to 200 patents in subclass 23.1 issuing per month.[111]

Table 1. The number of patents in Class 536 by subclass (23.1) or set of subclasses (23.1 to 23.7) for various two month intervals. Tallies for the two periods early in 2001 and 2000 are shown in parentheses.

YEAR	PERIOD	23.1	23.1 - .7
2001	6/1-8/1	425	233
2001	1/1-3/1	(333)	(197)
2000	6/1-8/1	351	186
2000	1/1-3/1	(396)	(188)
1999	6/1-8/1	329	154
1998	6/1-8/1	374	166
1997	6/1-8/1	235	91
1996	6/1-8/1	135	60

The USPTO declared in 1997 that it would issue patents on ESTs, and Incyte claims to have received the first EST patent in 1998 (Pat. No. 5,817,47)—although at least one patent that was issued in 1996 claims ESTs in addition to a full-length gene (Pat. No. 5,552,281). If patents in these subclasses prior to 1997 were not EST patents, then it is likely that a third of the patents in these classes after 1997 are not EST patents. If so, these numbers suggest that tens and perhaps a hundred EST patents issue every month.

I estimated references to various computational methods in EST patents as follows. I restricted my searches to patents in Class 536, Subclass 23.1 that claimed a polynucleotide sequence and included the terms "est" or "partial"

---

[109] Patent Number 5,817,479: Human kinase homologs (Oct. 6, 1998).

[110] Patent Number 6,194,152: Prostate tumor polynucleotide compositions and methods of detection thereof (February 27, 2001).

[111] The number of patents issuing could be limited by the number of examiners or the general availability of resources for examination of patents at the USPTO.

within 2 words of the terms (sequence or cDNA). I then searched for each of the following terms by monthly intervals: BLAST, Clustal (to indicate reference to the Clustal W method), Waterman (to indicate reference to the Smith-Waterman method), Markov (to indicate reference to a Hidden Markov Model), and GCG (to indicate the use of GCG software). I present the total number of patents in each category by year except for 2001; in 2001, I estimated the tally for the year by doubling the number of patents in each category for the period from January 1 to July 1.

Table 2. The number of likely EST patents per year that mentioned each of several methods of computational analysis. *Twice the number observed for the period January 1 to July 1.

	METHOD				
YEAR	BLAST	Waterman	Clustal	Markov	GCG
2001*	48	28	22	8	32
2000	53	23	20	15	43
1999	96	50	17	12	88
1998	41	19	0	0	35
1997	3	2	0	0	1
1996	0	1	1	0	1

The data clearly show that references to BLAST and Smith-Waterman began to be incorporated into patents issuing in 1998. The following year, patents began to issue that provided reference to Clustal W analysis and Markov Models. The number of references was similar for all methods in all remaining years, except that BLAST and Smith-Waterman methods were referenced about twice as many times in patents that issued in 1999 as in other years after 1997.

As shown in Table 1, the number of patents in Class 536, Subclass 23.1 did not change significantly from 1998 to 2001. The tallies for the number of patents in the restricted set used to examine the computational methods was not made, but is likely similar. Thus, the tallies shown here may estimate the frequency of mention of the various methods in patents in this restricted set of patents. However, the data suggest that the USPTO did begin issuing EST patents after it announced in 1998 that it would do so. Because this announcement came mid-year, the tallies for 1998 may underestimate the rate of mention of the methods in this year.

These preliminary data indicate that the USPTO began, in 1998, to issue a significant number of patents in Class 536, Subclass 23.1 that claimed nucleotide sequences, likely mentioned partial cDNA or EST, and referenced a method of sequence alignment. Furthermore, the USPTO has continued to issue such patents, at a seemingly similar rate, since 1998.

C. Satisfying the Written Description Requirement

The quantitative data suggest that the USPTO is issuing EST patents that rely on computational methods. To assess whether, and if so, how these or other computational methods are being used to address the written description requirement, I examined twenty-three patents and the USPTO Training Materials for evaluation of the written description.[112] I review the legal criteria and then assess the patents in light of the law.

1. Synopsis of Legal Criteria

An invention must be adequately described to qualify for a patent. The written description requirement is set forth in Section 112 of the Patent Act, its application to gene patents was addressed by the CAFC in several cases in the early 1990s, and the USPTO published guidelines in January 2001 explaining how to apply the requirement to various biotech claims, including claims to ESTs.[113] I briefly review the requirement here.

In general, the statute requires that the inventor describe the invention well enough to show "possession" of it. That is, the inventor must describe the invention in sufficient detail that a person "skilled in the art" would conclude the inventor actually invented the claimed invention.[114]

The CAFC determined in the early 1990s that, in order to describe a gene, an inventor must describe the DNA, purportedly in "structural" terms. For example, it is not enough to name the protein that the gene encodes and a method for isolating and sequencing the gene (even if it would be scientifically obvious how to isolate and sequence the gene). The inventor must give a "precise definition [of the DNA], such as by structure, formula, chemical name, or physical properties."[115] The rule was often (and inaccurately) simplified as requiring a description of the nucleotide sequence.

The CAFC acknowledged that a set of nucleotide sequences encoding a particular amino acid sequence could be deduced using the genetic code, but it emphasized the difference between deducing a set of possible sequences and knowing a naturally occurring sequence: If the amino acid sequence was newly discovered but the nucleotide sequence unknown, the inventor could claim only the set of all possible nucleotide sequences encoding it. But regardless whether the amino acid sequence for a protein was known or unknown, an inventor could discover and claim the nucleotide sequence that actually occurs in a particular organism.

The USPTO's Written Description Guidelines acknowledge these basic points. They emphasize, though, that "there is no basis for a per se rule requiring disclosure of complete DNA sequences or limiting DNA claims to only the

---

[112] U.S. Patent and Trademark Office, Synopsis of Application of Written Description Guidelines, *available at* http://wais.access.gpo.gov [hereinafter Written Description Training Materials].

[113] *See* Part II.B.1 further discusion of the CAFC's rulings and Part II.B.3 for further discussion of the USPTO's guidelines.

[114] *See, e.g,* Lockwood v. American Airlines, 107 F.3d 1565,1572 (Fed. Cir. 1997)

[115] Univ. of California v. Eli Lilly & Co., 119 F.3d 1559, 1556 (Fed. Cir. 1997).

sequence disclosed."[116] They therefore instruct examiners to consider "the level of skill and knowledge in the art, partial structure, physical and/or chemical properties, [and] *functional characteristics alone or coupled with a known or disclosed correlation between structure and function*" (emphasis added) in assessing the adequacy of a written description.[117]

2. Observed Uses of Computational Methods

An isolated DNA sequence that has utility may be claimed directly and is adequately described by giving its nucleotide sequence. Such a patented claim could easily be avoided, though, by changing a nucleotide so that the encoded amino acid sequence remains the same, or by changing an amino acid so that the function of the protein remains the same. Most inventors would like to state a claim that encompasses all these variants, and computational methods make that possible.

Computation methods cannot be used, though, to describe a set of nucleic acids that could vary in unpredictable ways. For example, a nucleotide sequence of a cDNA fragment or EST can often be shown by various sequence alignment methods to be homologous to a known DNA molecule that encodes a known protein of known function. However, if "gene" is defined to include naturally occurring regulatory elements and untranslated regions necessary and sufficient to mediate the expression of a cDNA, then the description of the cDNA fragment does not adequately describe the homologous gene. The USPTO Training Materials explain that the description is inadequate because "there is no known or disclosed correlation between th[e protein's] function and the structure of the non-described regulatory elements and untranslated regions of the gene."[118]

In short, computational methods can be used to describe a claimed set of nucleic acids when all the members of the set are expected to have the same function because of structural similarities. I found three methods for expanding the scope of a claim to a DNA sequence: by using the genetic code to define all the nucleic acids encoding the same polypeptide, by using percent identity to describe structurally similar sequences, and by identifying functional variants of particular amino acids. I discuss each in turn.

a. Use of the Genetic Code and Combinatorics

The most obvious way to define a set of nucleic acids that vary structurally but not functionally takes advantage of the degeneracy of the genetic code. Because there is more than one codon for many of the amino acids, there may be

---

[116] Written Description Guidelines, *supra* note 31, at 1101 ("Describing the complete chemical structure, i.e., the DNA sequence, of a claimed DNA is one method of satisfying the written description requirement, but it is not the only method.").

[117] Written Description Guidelines, *supra* note 31, at 1106

[118] Written Description Training Materials, *supra* note 112. Even if "gene" is not so defined, the description of a single cDNA is probably inadequate to claim all nucleic acids comprising it because it is not necessarily representative of that class; a "representative number" of such fragmnts are needed. *Id.* at 31-32.

a large number of nucleotide sequences that code for the same amino acid sequence. Defining that set of nucleotide sequences is a straightforward matter of mapping and combinatorics—even though there may be a very large number of nucleic acid sequences coding for a particular amino acid sequence (especially if the amino acid sequence comprises more than a few amino acids).

The USPTO Training Materials[119] acknowledge the reliability of this association between nucleotide structure and polypeptide structure. They explain that a claim to "[a]n isolated DNA that encodes protein X (SEQ ID NO: 2). ..."adequately describes a genus of molecules because "a person of skill in the art could readily envision all the DNAs degenerate to SEQ ID NO:1 by using a genetic code table" and "[o]ne of skill in the art would conclude that [the] applicant was in possession of the genus based on the specification and the general knowledge in the art concerning a genetic coding table."[120] Thus, the genetic code and combinatorial methods can be used to describe and claim the set of DNAs that encode a particular polypeptide.

The code is thus used to infer a set of nucleic acids encoding an experimentally determined amino acid sequence. For example, Incyte determined the amino acid sequence of a human ubiquitin-conjugating enzyme ("SEQ ID NO:2") and then patented the set of nucleic acids encoding that enzyme by claiming "[a]n isolated and purified polynucleotide encoding a polypeptide comprising an amino acid sequence of SEQ ID NO:2."[121] The code can also be use to infer the amino acid sequence from an experimentally determined nucleic acid sequence. For example, Incyte inferred the amino acid sequence of a protein it called "prostate expressed chemokine" from the cDNAs sequences it identified in a prostate cDNA library, and then claimed all the nucleic acids encoding that enzyme.[122]

All of the recently issuing patents that were assigned to DuPont or Incyte used this technique to claim a genus of DNAs encoding a given amino acid sequence. (XXX Add excerpts from patents explaining this.)

b. Use of Percent Sequence Identity

Perhaps the simplest way to define a set of similar amino acid sequences—or a set of nucleic acids encoding a set of similar amino acid sequences—relies on the similarity of their sequences to a described sequence. Such similarity is usually defined by the percentage of nucleic acids or amino acids that are identical ("percent identity") when a sequence in the set is aligned in some way with the described sequence.[123] The definition of a set of sequences by percent

---

[119] Written Description Training Materials, *supra* note 112, at 41-42.

[120] *Id.*

[121] Patent Number 6,277,568.

[122] Patent Number 5,840,544 (claiming "A purified polynucleotide encoding a polypeptide with an amino acid sequence shown in SEQ ID NO:2.").

[123] All of the EST patents assigned to DuPont noted simply that "[s]ubstantially similar nucleic acid fragments of the instant invention may also be characterized by the percent identity of the amino acid sequences that they encode to the amino acid sequences disclosed herein, as determined by algorithms commonly employed by those skilled in this art."

identity presumes the use of some method of sequence alignment, and the percent identity depends upon how the sequences are aligned.[124] If gaps are introduced to align the sequences, the corresponding amino acids or nucleic acids are typically ignored in calculating the percent identity.[125]

The USPTO Training Materials provide an example of the valid use of measures of percent identity to describe a set of proteins.[126] In the example, the inventor claims all variants of a protein having amino acid sequence X "that are at least 95% identical to X *and* catalyze the reaction of A B" (emphasis added).[127] This example thus alludes to a potential problem: Proteins that are at least 95% identical to X in structure might *not* be functionally similar. The example given addresses this problem by constraining the set of structurally similar proteins to those that are also functionally similar; it does not discuss any particular method alignment.

Alignment methods are used most simply to describe a set of nucleic acids that are similar to one or more specified nucleic acid sequences. For example, two patents issued recently to Bayer claim "[a]n isolated nucleic acid molecule consisting of a nucleotide sequence at least 85% identical to a sequence selected from the group consisting of SEQ ID Nos. [1, 2, . . . X]".[128] Alignment methods are also used to describe a set of polypeptides that are similar to one or more specified amino acid sequences. Those amino acid sequences may be deduced from an isolated nucleic acid sequence using the genetic code. For example, one of several EST patents issued to DuPont claims "[a]n isolated polynucleotide comprising . . . a nucleotide sequence encoding an isoleucyl-tRNA synthase, wherein the amino acid sequence of the synthase and the amino acid sequence of [sequence 2, 4, 6, or 8] have at least 80% identity based on the Clustal alignment method . . ."[129].

The method used to align the sequences is often but not always specified in the claims.[130] However, the minimum degree of similarity between the given

---

[124] The specification will usually describe at least one such method of sequence alignment. One patent noted several, including FASTA, BLAST, or ENTREZ (as part of the GCG package), Needleman and Wunsch, and Smith-Waterman methods. Patent Number 6,262,333.

[125] Percent Identity is defined in one patent as "the percentage of amino acid residues in a candidate sequence that are identical with the amino acid residues in the native sequence, after aligning the sequences and introducing gaps, if necessary, to achieve the maximum percent sequence identity, and not considering any conservative substitutions as part of the sequence identity," Patent Number 6,194,152. As explained in another patent, "[t]he percentage similarity between two amino acid sequences, e.g., sequence A and sequence B, is calculated by dividing the length of sequence A, minus the number of gap residues in sequence A, minus the number of gap residues in sequence B, into the sum of the residue matches between sequence A and sequence B, times one hundred. Gaps of low or of no similarity between the two amino acid sequences are not included in determining percentage similarity." Patent Number 6,277,568.

[126] Written Description Training Materials, *supra* note 112, at 54.

[127] *Id.*

[128] Patent Number 6,262,333 and Patent Number 6,262,234.

[129] Patent Number 6,271,441. Very similar claims are made in Patent Number 6,251,668 and Patent Number 6,255,090.

[130] Almost all of the DuPont patents specify the use of a Clustal alignment in the claims and do not describe any methods in the specification; others describe several in the specification but mention none in the claims. In contrast, a patent issued recently to Dendrion is probably unnecessarily

sequence and the sequences in the claimed set must be specified in the claims. This cutoff is clearly arbitrary. Requiring a higher degree of sequence identity means that the claimed sequences are less likely to differ functionally, all equal. Thus, it is common to see a series of claims that differ only in the minimum degree of similarity required. For example, the first claim requires only 80 or 85% sequence identity, a second claim requires 90% identity, and a third claim requires 95% identity. This strategy admits the possibility that a claim to sequences that are only 80% identical might be invalid.[131]

Claims to a nucleotide sequence "*encoding protein A*" that has "at least X% similarity to sequence S" were common in the surveyed patents. They are potentially problematic, though, because they do not explicitly require that the claimed structurally similar sequences have the same function as the isolated sequence or sequences.[132] Such functional similarity could be inferred by the reference to the protein by its name. However, several patents claimed a nucleotide sequence "encoding a protein *having the activity of protein A*" that has "at least X% similarity to sequence S."[133] They thereby restricting the claimed set of structurally similar nucleic acids to those that have a particular biochemical function.

c. Use of Structural Variants Having Similar Function

A more complex but potentially more accurate way to define a set of nucleic acids that vary structurally but not functionally considers the effect of amino acid substitutions on the structure and function of a molecule. Many amino acids may be replaced with other amino acids without changing the structure or function of the molecule. Information about the substitutability of amino acids can therefore be used to describe a set of nucleic acid sequences encoding a set of functionally similar polypeptides.

---

specific about the methods to be used when it claims "[a]n isolated polynucleotide having at least 95% sequence identity to nucleotides 43-3327 of the sequence of SEQ ID NO: 14, wherein % identity is calculated using the LALIGN program found in the FASTA Version 2.0 suit of programs using default parameters with the BLOSUM50 matrix, a ktup of 2 and a gap penalty of -12/-2 . . ." Patent 6,194,152

[131] A patent typically has many claims, which vary in scope from very broad to very narrow. The broadest claims are most likely to be found invalid by a court, but the narrowest claims are unlikely to be infringed because they are easy to work around. The use of a series of claims of decreasing scope is a strategy to ensure the broadest possible valid claim. This strategy was used in most of the DuPont patents that I read; it was not used, for example, in Patent Number 6,242,256.

[132] Compare such a claim to Example 14 in the Written Description Training Materials, *supra* note 112; *see also* text accompanying note 123.

[133] For example, Patent Number 6,262,345 claimed "[a]n isolated polynucleotide comprising . . . a nucleotide sequence *encoding a polypeptide having glycogen synthase kinase activity*, . . . wherein the amino acid sequence of the polypeptide and the amino acid sequence of [sequence 1, 2, . . . X] have at least 90% identity based on the Clustal alignment method. . .". A claim in Patent Number 6,274,379 is similar. Patent Number 6,277,568 claimed "[a]n isolated and purified polynucleotide having at least 90% sequence identity . . . [to] the polypeptide of SEQ ID NO: 2, and which *encodes a polypeptide that retains ubiquitin-conjugating activity.*").

Amino acids may differ, for example, "in polarity, charge, solubility, hydrophobicity, hydrophilicity, and/or the amphipathic nature of the residues."[134] If the substituted amino acid has similar characteristics, the change is "conservative" and is unlikely to change the structure or function of the protein. Substitutions involving amino acids with very different attributes are "non-conservative" and may produce "[s]ubstantial changes in function or immunological identity. . . For example, substitutions may be made which more significantly affect the structure of the polypeptide backbone in the area of the alteration, for example the alpha-helical or beta-sheet structure, the charge or hydrophobicity of the molecule at the target site, or the bulk of the side chain."[135]

The USPTO training materials do not discuss the use of methods of amino acid substitution to describe a genus of nucleic acid. But many inventors discuss "variants" of a polypeptide in the specification of the patent made by either conservative or non-conservative substitutions.[136] They often indicate that conservative variants are within the scope of the claimed invention[137] and may specify methods for determining conservative substitutions.[138]

D. Satisfying the Utility Requirement

The USPTO is clearly issuing patents that rely on computational methods to *describe* a genus or set of nucleic acid sequences. To assess whether, and if so, how computational methods are being used to establish the *utility* of patents for partial cDNAs or ESTs, I examined the USPTO Utility Training Materials[139] and the same twenty-three patents that I used to assess whether and how computational methods are being used to address the Written Description Requirement. As in the last section, I review the legal criteria and then assess the patents in light of the law.

1. Synopsis of Legal Criteria

An invention must be useful to qualify for a patent. The utility requirement is set forth in Section 101 of the Patent Act, but its application to gene patents has

---

[134] Patent Number 6,277,568.

[135] Patent Number 6,194,152.

[136] A "variant" . . .may have an amino acid sequence that is different by one or more amino acid "substitutions". The variant may have "conservative" changes, wherein a substituted amino acid has similar structural or chemical properties, e.g., replacement of leucine with isoleucine. More rarely, a variant may have "nonconservative" changes, e.g., replacement of a glycine with a tryptophan." Patent Number 5, 840,544.

[137] For example, one inventor indicate that "[d]eliberate amino acid substitutions may be made on the basis of similarity in polarity, charge, solubility, hydrophobicity, hydrophilicity, and/or the amphipathic nature of the residues, as long as the biological or immunological activity of EXADH is retained." Patent Number 20010010913.

[138] *E.g.* "Guidance in determining which and how many amino acid residues may be substituted, inserted or deleted without abolishing biological or immunological activity may be found using computer programs well known in the art, for example, DNASTAR software." Patent Number 5,840,544.

[139] Utility Training Materials, *supra* note 97.

not yet been addressed by the CAFC; nonetheless, the USPTO published guidelines in January 2001 explaining how to apply the requirement to various claims in biotech patents, including patents on ESTs.[140] I briefly review the requirement here.

The Supreme Court held in 1966 that an invention must have a real world, practical utility.[141] In that case, it found that a process for making a chemical that was used only in research lacked such utility.[142] Various appellate court cases since then have held, in addition, that an invention must have a "specific and substantial" utility.[143] And prior to 1996, the USPTO required its examiners to determine whether an invention had a "credible" or well-established utility.

The USPTO's new Utility Guidelines require that all claimed inventions have a "specific, substantial, and credible utility."[144] A "credible" utility is logically consistent with the asserted facts. For example, since at least some nucleic acids can be used as probes or chromosome markers, it is credible that any particular DNA can be used in this way. A "substantial" utility is a real-world use. For example, a claim that a nucleic acid is useful as a dietary protein supplement is insufficient; it is a "throw-away" use that lacks substance. A "specific" utility is particular to the subject matter claimed. For example, if a nucleic acid is claimed to be useful as a gene probe or chromosomal marker, then the specific DNA target must be disclosed.

The new Utility Guidelines raised the bar on utility because inventions must now have a substantial and specific use—not just a credible one. However, the procedural requirements for evaluating utility clearly favor the patent applicant. USPTO personnel must presume that statements by applicants are true, and they must allow applicants to rebut any prima facie finding of no utility.[145]

Despite the wishes of many commentators, the new Utility Guidelines do not create a create a "per se" rule against homology-based assertions of utility. The PTO said there is no "scientific evidence that homology-based assertions of utility are inherently unbelievable or involve implausible scientific principles.[146] Instead of an across-the-board rule, the PTO declared that assessments of utility would be "fact dependent" and determinations would be made "on the basis of scientific evidence."[147]

2. Observed Uses of Computational Methods

---

[140] *See* Part II.B.1 further discusion of the CAFC's rulings and Part II.B.3 for further discussion of the USPTO's guidelines.

[141] 383 U.S. 519 (1966).

[142] 383 U.S. 519 (1966).

[143] need to get some examples or summary citations here.

[144] The guidelines also discuss a "well-established" utility test, but even well-established utilities must be specific, substantial, and credible. However, if the utility is well-established, it need not be asserted explicitly in the patent. For an excellent review and critique of the Utility Guidelines, see Worrall, *supra* note 96, 132.

[145] Utility Training Materials, *supra* note 97; Worrall, *supra* note 96, at 132.

[146] Utility Guidelines, *supra* note 21, at 1096.

[147] *Id.*

Claims to nucleic acid sequences as compositions of matter must assert a credible and specific practical utility for the sequence. A nucleic acid may be useful because it encodes a particular known and useful protein, or because it can be used as a probe to identify or locate the full-length nucleic acid encoding a specific known and useful protein. Even if the function of the encoded protein is unknown, a nucleic acid that is transcribed in some cells but not others may be useful as a diagnostic tool—if its presence is correlated, for example, with a particular disease.[148]

The utility of a nucleic acid thus often (but not always!) requires information about the biological function of the particular encoded polypeptide.[149] Such information may be obtained directly and experimentally in the laboratory. It may also be inferred from comparison to sequences whose function has already been directly and experimentally determined in the laboratory. The latter technique requires computational methods of sequence alignment and is the more contentious method for establishing the utility of a sequence.[150]

In short and despite the debate, computational methods may be used to establish the utility of ESTs by comparing the partial or complete cDNA sequences to full length sequences encoding proteins of known function, and then inferring the function of the protein partially or completely encoded by the cDNA sequence.[151] The patents that I examined used computation methods in precisely this fashion.

a. To Identify the Polypeptide Encoded by a Sequence

---

[148] "[T]he utility of a claimed DNA does not necessarily depend on the function of the encoded gene product. A claimed DNA may have a specific and substantial utility because, e.g., it hybridizes near a disease-associated gene or it has a gene-regulating activity." Utility Guidelines, *supra* note 21, at 1095.

[149] As demonstrated in Example 9 of the Training Materials, a set of cDNAs is not useful merely because they encode part of some protein and can be used a probes to identify the full length nucleic acid encoding that protein; the particular protein that they encode must be determined and specified. Utility Training Materials, *supra* note 97, at 50-53. However, establishing the function of the encoded polypeptide is only one way to establish real world utility. Real-world utility and the function of the gene are frequently but inaccurately treated as synonyms. For example, the statement that "[p]atent applications that do not specify exactly what a gene or gene fragment is, or what its function is, will not be considered for approval, according to the new guidelines" confuse real-world utility and gene function. *Updated Guidelines from Patent Office Similar to Old Ones*, BIOTECHNOLOGY NEWSWATCH at 9 (Feb. 5, 2001).

[150] Experimental evidence is typically considered more reliable than the "hypotheses" or "theoretical results" resulting from the analysis of genomic databases. For example, one author noted that "[o]pen reading frames vary widely in the degree to which their encoded proteins assert a credible specific and substantial utility" and then explained that "[a]t one extreme, DNA sequences encoding proteins having experimentally verified function and use satisfy the utility requirement. At the other extreme, the function of an unknown protein can be *hypothesized* based on sequence similarity, or homology, to known sequences with known function." Worrall, *supra* note 96, at 139. *See also* notes 81-83, *infra*, and accompanying text.

[151] "[W]hen a patent application claiming a nucleic acid asserts a specific, substantial, and credible utility, and bases the assertion upon homology to existing nucleic acids or proteins having an accepted utility, the asserted utility must be accepted by the examiner unless the Office has sufficient evidence or sound scientific reasoning to rebut such an assertion".Utility Guidelines, *supra* note 21, at 1096.

The USPTO training materials provide an example of the use of computational methods to assess the structure and function of the protein encoded by a full open reading frame, and thereby satisfy the utility requirement.[152] In the example, a cDNA library is prepared, clones are sequenced, and their open reading frames are identified. The nucleic acid sequence is found to be similar to various known ligases, presumably by doing sequence alignments. The amino acid sequence that it encodes is compared to a consensus sequence of the known ligases, and "reveals a similarity score of 95%." The nucleic acid sequence also has a "high homology" to DNA Ligase encoding nucleic acids, and has only 50% "homology" to the next most similar sequence. The Training Materials indicate that these disclosures are sufficient to establish that the claimed sequence encodes a DNA ligase and, since DNA ligases have "a well-established use in the molecular biology art," the disclosure establishes a utility for the claimed sequence.

This basic method was used in the patent that Incyte claims was the first EST patent to issue.[153] The patent describes 44 partial cDNAs that were isolated from various cDNA libraries. According to the specification, each nucleotide and its corresponding amino acid sequence was compared to sequences in GenBank using a proprietary search algorithm,[154] and homologous regions were identified. The specification does not provide any statistics or results from the analysis of the described sequences.[155] It does, however, note that "protein kinases are associated with basic cellular processes such as cell proliferation, differentiation and cell signaling" and asserts that "[k]inase nucleotide sequences are [therefore] useful in diagnostic assays used to evaluate the role of a specific kinase in normal, diseased, or therapeutically treated cells". A patent issued soon thereafter is very similar.[156]

The same approach was used in two recently issued patents that were assigned to Incyte. In a patent on human ubiquitin-conjugating enzymes, Incyte took clones from a prostate cDNA library and then used BLAST to ascertain that one of them had "chemical and structural similarity with *Arabidopsis thaliana* [a plant] ubiquitin-conjugating enzyme (GI 1707021)."[157] The threshold for the BLAST was given as 10^{-25} for nucleotides and 10^{-8} for polypeptides. Thus, the probability that the newly discovered polypeptide sequence was the same as the *Arabidopsis* gene purely by chance was less than 10^{-8}, and the new sequence was inferred to be a ubiquitin-conjugating enzyme. Because ubiquitin is part of a pathway for

---

[152] Utility Training Materials, *supra* note 97, at 53-55.

[153] Patent Number 5,817,479

[154] The search algorithm was "developed by Applied Biosystems and incorporated into the INHERIT TM 670 Sequence Analysis System" and used "Pattern Specification Language (TRW Inc, Los Angeles, Calif.)" to determine regions of homology.

[155] It merely explains in general that dot matix plots were used "to distinguish regions of homology from chance matches" and Smith-Waterman alignments were used "to display the results of the homology search". The specification also explains that BLAST could also be used to find High-scoring Segment Pairs, whose probability score meets a predetermined threshold level of significance.

[156] Patent Number 5,840,544.

[157] Patent Number 6,277,568.

selective protein degradation, the claimed sequence was asserted to be "useful in the diagnosis, treatment, and prevention of cancer, autoimmune disorders, and neuronal disorders."

In a second patent issuing recently and assigned to Incyte, this one on human extracellular adhesive proteins,[158] Incyte used more seemingly more comprehensive but only vaguely described methods to identify homologs of known function.[159] And the specification merely asserts that the sequences are useful for diagnosing, treating or preventing disorders associated with expression of the proteins.[160]

Eleven EST patents that recently issued and were assigned to DuPont were very similar in structure and approach.[161] They all claimed partial cDNAs from plants, and the functions of the claimed sequences were usually determined by finding homologs of known function from humans or other animals. Most of them relied on BLAST to compare the isolated cDNAs to sequences in various databases.[162] And most of them included p-values for the comparison of each described sequence and the homolog used to infer its function, as well as the percent identity of the described sequence and its homolog.[163] P-values were typically smaller than 10^{-25}, and the claimed cDNAs were usually more than 70% similar to the sequences of known function.[164]

The patents assigned to DuPont were clearly distinguishable from the patents assigned to Incyte. They relied fairly exclusively on the described findings of

---

[158] Patent Number 20010010913.

[159] "The polynucleotide sequences were validated by removing vector, linker, and polyA sequences and by masking ambiguous bases, using algorithms and programs based on BLAST, dynamic programing, and dinucleotide nearest neighbor analysis. The sequences were then queried against a selection of public databases such as GenBank primate, rodent, mammalian, vertebrate, and eukaryote databases, and BLOCKS to acquire annotation, using programs based on BLAST, FASTA, and BLIMPS. The sequences were assembled into full length polynucleotide sequences using programs based on Phred, Phrap, and Consed, and were screened for open reading frames using programs based on GeneMark, BLAST, and FASTA. The full length polynucleotide sequences were translated to derive the corresponding full length amino acid sequences, and these full length sequences were subsequently analyzed by querying against databases such as the GenBank databases (described above), SwissProt, BLOCKS, PRINTS, PFAM, and Prosite." Patent Number 20010010913.

[160] However, it recites a list of potentially treatable diseases that is 31 lines (>300 words) long!

[161] *See* note 107, *infra.*

[162] Typically, the cDNA sequences were "analyzed for similarity to all publicly available DNA sequences contained in the "nr" database using the BLASTN algorithm, [and] . . . [t]he DNA sequences were translated in all reading frames and compared for similarity to all publicly available protein sequences contained in the "nr" database using the BLASTX algorithm . . . ." Patent Number 6,255,090. Slightly different language is used in Patent Number 6,255,114.

[163] The specifications explained that "the P-value[s] (probability) of observing a match of a cDNA sequence to a sequence contained in the searched databases merely by chance as calculated by BLAST are reported herein as "pLog" values, which represent the negative of the logarithm of the reported P-value. Accordingly, the greater the pLog value, the greater the likelihood that the cDNA sequence and the BLAST "hit" represent homologous proteins."

[164] In two patents claiming transcription coactivators from plants by homology to mouse and human proteins, the claimed sequences were only 19-46% identical to the sequences of known function; the p-values were all less than 10^{-20}. Patent Number 6,255,090 and Patent Number 6,271,441.

homology to establish utility; that is, they typically did not include any additional laboratory work on the claimed sequences. The asserted utility of the cDNAs claimed in the DuPont patents also tended to be less explicit and more general in nature than the utility asserted in the Incyte patents. In general, the DuPont patents relied on sequence comparisons to claim sequences identified in an early stage of research, whereas the Incyte patents used sequence comparisons in combination with a variety of laboratory findings to justify their claims to such sequences.

b. To Show that the Polypeptide is Unknown

An expressed nucleic acid sequence may be useful even if the biological function of the protein that it encodes is unknown. The utility arises from knowledge of factors that are correlated with the expression of the sequence. For example, many sequences are expressed only in cancerous cells; these sequences are therefore useful as indicators of cancer. Several recently issued EST patents use computational methods to demonstrate that sequences are novel, and then assert utility based on their specificity to particular types of tumor or cancer cells.[165]

Computational methods may also be used to establish whether or not a sequence is known or has known homologs so that research and patents can be designed appropriately. For example, if there are no known homologs of a sequence, its function cannot be inferred from the analysis of genomic databases but additional research may be advantageous. If the exact sequence is already described, additional research is unnecessary and the sequence itself cannot be claimed. However, it is possible that the sequence can be claimed as an indicator of disease.[166]

E. Discussion and Critique

Randall Scott of Inctye asserts that "there are many, many families [of genes] now for which the function can be reasonably predicted from the structure, and [our ability to predict function from structure gets] better and better . . . every year.[167] He was presenting testimony to a Congressional Hearing on Genomic Inventions, arguing for the patentability of ESTs whose utility was established

---

[165] For example, computational methods were used to establish that sequences specific to human prostate tumor cells were novel. Patent Number 6,194,152.

[166] A patent assigned to Incyte for concensus sequences from cancer cells reports whether or not each sequence has a known homolog; if it does, then the specification adds that the sequence has now been observed from a cancer cell. Similarily, two patents assigned to Bayer are careful to distinguish "1) matches to known human genes, 2) matches to human EST sequences, and 3) no significant match to either 1 or 2, and therefore a potentially novel human sequence." Patent Number 6,262,333 and Patent Number 6,262,334.
Patent Number 5,932,442.

[167] Dr. Randal W. Scott, President And Chief Scientific Officer, Incyte Genomics. Prepared Statement at Congressional Hearing on Genomic Inventions, *supra* note 1.

from comparison to sequences of known function. His comments reflect both legal and scientific problems in inferring function from structure.

Patent law requires that every invention be adequately described and have a practical utility. The courts have made it clear that a nucleotide sequence can only be patented when its "structure" is adequately described[168]. However, it must also have an asserted utility, which is often only possible when the function of the encoded protein is know. Thus, to patent a gene sequence or set of gene sequences, one must usually know both its structure and the function of the encoded protein or proteins.

Discussions of about the patentability of genes, especially partial cDNAs or ESTs, have tended to focus on the utility requirement. However, the utility requirement and the written description requirement are flip sides of same coin, because both create issues about the use of computational methods to translate between structure and function.

The genetic code provides one biological reality that has required an adjustment to the idea that a nucleic acid must be structurally described in order to be patented. It allows one structure (i.e. an amino acid sequence) to be reliably translated into another (i.e. a nucleotide sequence), and vice versa if the reading frame is known. The legal world struggled to distinguish a claim to a "theoretical" genus of nucleotide sequences from a claim to a naturally occurring nucleotide sequence, but the basic idea is simple and sound. All the nucleic acids that encode a polypeptide can be patented if the amino acid sequence of the polypeptide is known, because all those sequences code for the same polypeptide structure.

The description of all the nucleotides that encode a set of "similar" amino acid sequences (or a set of "similar" nucleotides) by measures of percent identity is more problematic because structural similarity does not correlate exactly with functional similarity. Some differences in some amino acids are more important than others. Definition of sequences by their percent similarity is computationally simple and it provides a bright-line test for deciding whether two sequences are similar or not. However, unless the definition is extremely rigid, so that only very similar sequences are considered the same, it will probably include sequences that encode polypeptides with other functions—however slightly.

Ideally the definition of a set of sequences will clearly distinguish those that are functionally similar and those that are not, and that threshold can be accurately determined. In other words, the receiver operator curve for the method will have a sharp transition, indicating a clear separation between true positives and false positives.

The USPTO addresses this problem of identifying functionally similar sequences by proposing the definition of a set of nucleotides that share some degree of structural similarity *and have the same activity as the given sequence.* However, this technique poses legal problems because the CAFC seems to have asserted that functional attributes cannot be used to define a claimed structure. The court could distinguish this technique by noting that it merely limits a set of

---

[168] Case law forbids the use of functional attributes to describe a claimed composition. *See* Part II.B.1.

structurally similar sequences, but that seems to push the structural definition rule beyond the bounds of legal or scientific reason.

It is, however, well-known in the art that some methods for comparing sequences or defining sets of sequences are better than others in identifying sequences of similar function based upon their sequence similarity. For example, a gapped Blast may provide a more functionally accurate analysis of sequence similarity than an ungapped Blast if there are many insertions or deletions. A multiple alignment that uses an appropriately selected substitution matrix results in fewer false positives than one that assumes all substitutions are equally likely. Hidden Markov Models may provide a better model for identifying functionally similar sequences than, for example, a simple gapped BLAST search.

It is also well known in the art that more exhaustive, more sensitive methods tend to be slower and are often more complex than others. In many cases, a simple, approximate method is sufficient to identify all functionally similar sequences; in other cases, it may not. The sufficiency of a method for assessing the similarity of sequences or defining a set of sequences will be case-specific, depending on the actual sequence landscape and the extent of clustering within that landscape. All else equal, simpler methods are probably preferable.

Patent applicants have addressed the problem of identifying functionally similar sequences by discussing the difference between conservative and non-conservative changes and appealing to the knowledge of one skilled in the art. This approach may avoid the legal problem of using function to define a structure, since the approach is based on inferences of functional equivalency of parts of the polypeptide rather than functional equivalency of the entire protein. It is philosophical related to the use of substitution matrices, but more flexible.

In short, the use of any method for assessing sequence similarity is potentially problematic when the measure of similarity is used to infer function. Methods that account for the greater likelihood of particular amino acid substitutions assume that such changes will not affect the proteins function, and may permit more accuate inferences of function from structure. Percent identity for sequences aligned with a model that uses reasonable parameters is probably a good and simply rule of thumb for describing a set of sequences that are likely to have similar function. The adequacy of the threshold may vary with the protein, though; for example, stricter thresholds may be necessary when function varies greatly with small changes in structure. Similarly, methods that assess the probability that a sequence is structurally similar to a protein of known function can probably often be used reliably, especially when the sequences are very similar.

V. CONCLUSION

The USPTO is issuing large number of patents on ESTs whose utility is often established by comparing them to sequences of known function, and allowing claims to sequences that share some critical but arbitrary percentage of identical nucleotides or amino acids. The methods used to infer utility and describe a

claimed set of sequences appear scientifically sound and will likely produce reliable results in most cases. Sequences that encode proteins with different functions are best excluded by reference to their difference in function.

The recent guidelines issued by the USPTO have clarified their position with respect to a number of issues: "A DNA sequence per se is not patentable. Isolated genes can be patented. The entire gene sequence doesn't have to be disclosed. The gene must have a use. An EST must have a use. The applicant only has to disclose one use for the gene. The gene's function doesn't have to be known in order for the DNA to be useful."[169] The guidelines are a declaration that "the patenting of genomic inventions is consistent with our law and with our practice."[170]

However, the CAFC has not ruled on either of the two more contentious issues involving the use of computational methods in describing partial cDNAs and identifying their utility. It is possible that the court will view these issues quite differently than the USPTO and scientists. The simultaneous failure of politicians to appreciate the scientific validity of genomic methods and the sophistication of patent applications with claims to ESTs is remarkable.

The business world is likely to have more effect on the issuance of patents than the courts and the USPTO. The USPTO says that it is seeing more "generation three" EST patents—patents whose utility is supported by more than "mere homology," and fewer "generation two" EST patents, whose utility is supported only by homology. However, my reading of several recent patents suggests that there are differences in the patenting strategies of companies in the human gene business and companies in the plant gene business. These strategies may reflect differences in publicity and political pressure.

In sum, the use of computational methods to identify the utility of ESTs and describes claims to similar sequences is probably scientifically and legally feasible—although not without problems on either account. How the issuance of such patent affects the progress of research and the development of industries that rely on genetic information is another issue.

---

[169] Van Brunt, *supra* note 24.

[170] Todd Dickinson. Statement at Congressional Hearing on Genomic Inventions, *supra* note 1.

Short blocks from the noncoding parts of the human genome have instances within nearly all known genes and relate to biological processes

Isidore Rigoutsos*, Tien Huynh, Kevin Miranda, Aristotelis Tsirigos, Alice McHardy, and Daniel Platt

IBM Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598

Communicated by Thomas E. Shenk, Princeton University, Princeton, NJ, March 4, 2006 (received for review November 16, 2005)

Using an unsupervised pattern-discovery method, we processed the human intergenic and intronic regions and catalogued all variable-length patterns with identically conserved copies and multiplicities above what is expected by chance. Among the millions of discovered patterns, we found a subset of 127,998 patterns, termed pyknons, which have additional nonoverlapping instances in the untranslated and protein-coding regions of 30,675 transcripts from 20,059 human genes. The pyknons arrange combinatorially in the untranslated and coding regions of numerous human genes where they form mosaics. Consecutive instances of pyknons in these regions show a strong bias in their relative placement, favoring distances of ≈22 nucleotides. We also found pyknons to be enriched in a statistically significant manner in genes involved in specific processes, e.g., cell communication, transcription, regulation of transcription, signaling, transport, etc. For ≈1/3 of the pyknons, the intergenic/intronic instances of their reverse complement lie within 380,084 nonoverlapping regions, typically 60–80 nucleotides long, which are predicted to form double-stranded, energetically stable, hairpin-shaped RNA secondary structures; additionally, the pyknons subsume ≈40% of the known microRNA sequences, thus suggesting a possible link with post-transcriptional gene silencing and RNA interference. Cross-genome comparisons reveal that many of the pyknons have instances in the 3′ UTRs of genes from other vertebrates and invertebrates where they are overrepresented in similar biological processes, as in the human genome. These unexpected findings suggest potential unique functional connections between the coding and noncoding parts of the human genome.

junk DNA | pattern discovery | posttranscriptional gene silencing | pyknons | RNA interference

The intergenic and intronic regions comprise most of the genomic sequence of higher organisms. Even though recent work suggested their participation in a regulatory role (1, 2), the true function of these regions remains largely elusive. The search for conserved motifs, presumed to be regulatory and control signals, upstream of the 5′ UTRs of genes has been the focus of research activities for many years (3–7).

Recently, researchers began studying the 3′ UTRs of genes where they discovered functionally significant conserved regions, in direct analogy to the cis-motifs of promoter regions (8). Comparative analyses permitted the study of conservation in the vicinity of genes and elsewhere in the genome (9–13) but were carried out on only a handful of organisms at a time because of the magnitude of the involved computations (14–17).

The analysis of 3′ UTRs intensified after they were discovered to contain binding sites that are targeted by short interfering RNAs and result in the posttranscriptional control of the corresponding gene's expression through either mRNA degradation or translational inhibition (18–27). Accumulating evidence that noncoding RNAs control developmental and physiological processes (28–32) and that a considerable part of the human genome is transcribed (33) led researchers to identify functional elements (34) in areas of the genome that are not associated with protein-coding regions.

Here, we examine whether highly specific patterns exist within a single genome that may act as targets or sources for putative regulatory activity or as a "vocabulary" for as yet undiscovered mechanisms. Our analysis represents a substantial point of departure from previous efforts. First, we carry out all of the analysis on a single genome. Second, we seek patterns in the intergenic and intronic regions of the genome (not the UTRs or protein coding regions). Third, our patterns transcend chromosomal boundaries. And fourth, we rely on the unsupervised discovery of recurrent variable-length sequence fragments instead of using searching schemes. We discovered >66 million motifs with multiplicities well above what is expected by chance. A sizeable subset of these motifs, referred to as the pyknons,† have one or more additional instances in the UTRs and coding regions (CRs) of almost all known human genes and exhibit properties that suggest a possibly extensive link between the genome's nongenic and genic regions and a connection with posttranscriptional gene silencing (PTGS) and RNA interference (RNAi).

Results

Pattern-Discovery Step. Using a version of a pattern-discovery algorithm we developed earlier (35), modified to handle very large data inputs, we sought variable-length motifs that are identically conserved across all of their instances, comprise a minimum of $L = 16$ nucleotides, and appear a minimum of $K = 40$ times in the processed input (see *Supporting Text*, which is published as supporting information on the PNAS web site, regarding the values of L and K). The algorithm guarantees the reporting of all composition-maximal and length-maximal patterns satisfying these parameters (see *Supporting Text*). The input comprised the intergenic and intronic sequences of the human genome from ENSEMBL Rel. 31 (36) and totaled 6,039,720,050 nucleotides. The input did not include the reverse complement of the 5′ UTRs, amino acid coding, or 3′ UTRs of any human genes. This exclusion ensures that any discovered patterns are not connected to the sequences of known genes, protein motifs, or domains (see *Supporting Text* for details). This step generated an initial set P_{init} of 66+ million, variable-length statistically significant patterns (see *Methods*). The *Supporting Text* contains information on the properties of P_{init}'s entries.

Notation/Convention. We will use CRs to refer to the translated, amino acid coding part of exons and also associate the colors blue, red, and yellow with 5′ UTRs, CRs, and 3′ UTRs, respectively.

Determining Which of the Discovered Patterns Have Additional Instances in the 5′ UTRs, CRs, or 3′ UTRs of Known Genes. We considered the members of P_{init} in order of decreasing value of the product

Conflict of interest statement: No conflicts declared.

Freely available online through the PNAS open access option.

Abbreviations: CR, coding region; PTGS, posttranscriptional gene silencing; RNAi, RNA interference.

*To whom correspondence should be addressed. E-mail: rigoutso@us.ibm.com.

†From the Greek adjective πυκνός/πυκνή/πυκνόν meaning "serried, dense, frequent."

(length-of-pattern × copy-number-of-pattern), ensuring that longer and more frequent patterns are considered before shorter and less frequent ones. We kept a pattern p only if none of its untranslated/CR instances collided with a previously kept pattern (see *Supporting Text*). After filtering kept patterns for low-complexity with NSEG (37), we generated three pattern sets $P_{5'\text{UTR}}$, P_{CR} and $P_{3'\text{UTR}}$ that contained 12,267, 54,396, and 67,544 patterns, respectively, and had one or more instances in 5′ UTRs, CRs, or 3′ UTRs. $P_{5'\text{UTR}} \cup P_{\text{CR}} \cup P_{3'\text{UTR}}$ contained 127,998 patterns, indicating that the three pattern sets are largely disjoint. We refer to these 127,998 patterns as pyknons. See *Supporting Text* for information on the sets $P_{5'\text{UTR}}$, P_{CR}, and $P_{3'\text{UTR}}$.

The pyknons exhibit a number of properties that connect the nongenic and genic regions of the human genome in unexpected ways, in particular, as discussed below.

The pyknons have one or more instances within nearly all known genes. The 127,998 pyknons that we originally discovered in the human intergenic and intronic regions have an additional 226,874 nonoverlapping copies in the 5′ UTRs, CRs, or 3′ UTRs of 20,059 genes (30,675 transcripts). That is, >90% of all human genes contain one or more pyknon instances. The pyknons in $P_{5'\text{UTR}}$ cover 3.82% of the 6,947,437 nucleotides in human 5′ UTRs; the pyknons in P_{CR} cover 3.04% of the 50,737,024 nucleotides in human CRs; and the pyknons in $P_{3'\text{UTR}}$ cover 7.33% of the 25,597,040 nucleotides in human 3′ UTRs.

The pyknons arrange combinatorially in many human 5′ UTRs, CRs, and 3′ UTRs, forming mosaics. The number of pyknon instances in human transcripts is skewed (see *Supporting Text*). More than 16,000 transcripts contain at least 4, whereas ≈2,200 transcripts contain 20 or more pyknon instances in their UTRs and CRs. In those cases where we find many pyknons, they arrange combinatorially and form mosaics. Fig. 1 shows an example of such a combinatorial arrangement in the 3′ UTRs of *birc4* (an apoptosis inhibitor) and nine other human genes. The 3′ UTR of *birc4* contains 100 instances of 95 distinct pyknons; of these, 22 are also present in the 3′ UTRs of the other nine genes shown. One or more instances of the 95 pyknons from *birc4*'s 3′ UTR exist in the 3′ UTRs of 2,306 transcripts (data not shown). The *Supporting Text* includes examples of similar combinatorial arrangements of pyknons in the 5′ UTRs and CRs of known genes. Recall that we initially discovered the pyknons in an input that included neither transcribed gene-related sequences nor their reverse complement.

The pyknons account for 1/6 of the human intergenic and intronic regions. The intergenic and intronic copies of the pyknons span 692,393,548 positions on the forward and reverse strands. For those pyknons whose reverse complements are not already in the list of 127,998 pyknons, their Watson-strand instances impose constraints on their Crick-strand instances. Taking this observation into account and recalculating shows that pyknons and their reverse complement cover 898,424,004 positions or ≈1/6 of the human intergenic/intronic regions.

The pyknons are nonredundant. We clustered the pyknons using a BLASTN-based scheme (38). Because our collection includes pyknon pairs whose members are the reverse complement of one another, we had to ensure that the clustering scheme did not overcount: when comparing sequences A and B, we examined for redundancy the pair (A,B) and the pair (reverse-complement-of-A,B). Clustering at X = 70%, 80%, and 90%, we generated clusters with 32,621, 44,417, and 89,159 pyknons, respectively (see *Supporting Text* for details). The high numbers of surviving clusters suggest that the pyknons are largely distinct.

On pyknons and repeat elements. One thousand two hundred ninety-two pyknons (1.0%) have instances occurring exclusively inside repeat elements, as determined with the help of REPEATMASKER (Smit, A. & Green, P. RepeatMasker: http://ftp.genome.washington.edu). Seventy-nine pyknons have instances exclusively in repeat-free regions. The remaining 126,627 pyknons (98.9% of total) have instances both inside repeat elements and in repeat-free regions. See *Supporting Text* for details.

The pyknons are distinct from the "ultraconserved elements." Fifty-two pyknons have instances in 46 of the 481 ultraconserved elements (9) and cover 0.67% of the 126,007 positions: uc.73+ contains four pyknons; uc.23+, uc.66+, uc.143+, and uc.414+ each contain two pyknons; the remaining 41 elements contain a single pyknon each.

The pyknons are associated with specific biological processes. For 663 Gene Ontology (GO) terms (39) describing biological processes at varying levels of detail, we found that the corresponding genes had either a significant enrichment or a significant depletion in pyknon instances; Table 1 shows a partial list of GO terms that are enriched or depleted in pyknons. The full list appears in Table 4, which is published as supporting information on the PNAS web site.

The relative positioning of pyknons in 5′ UTRs, CRs, and 3′ UTRs is strongly biased, but consecutive pyknon instances are not correlated. We examined the distances between consecutive pyknons, separately for the 5′ UTRs, CRs, and 3′ UTRs: Fig. 2 shows the calculated probability density functions. The curves have similar shapes, pronounced peaks at abscissas 18 and 22, and a preference for distances between 18 and 31 nucleotides, suggesting a tight packing of pyknons in these regions that favors the distances shown in the histogram. We considered the possibility that the pyknon instances are fragments of larger regions that are conserved in genic and nongenic regions. Let b be a pyknon instance in 5′ UTR, CR, or 3′ UTR, and let us assume that, unknown to us, b is part of a larger-size conserved unit B. Then B will span an area larger than is delineated by b, and there will be $length(B) - length(b) + 1$ strings in the immediate neighborhood of b that would have as many identically conserved intergenic and intronic copies as b. We checked this in 3′ UTRs by taking each instance of a pyknon in $P_{3'\text{UTR}}$, shifting it by $+d$ (respectively $-d$), generating a new string and locating the new string's instances in the human intergenic and intronic regions. Had the pyknons been part of larger conserved units, then for some values of d, the number of intergenic and intronic copies of the newly formed strings would have remained identical to those of the starting strings. On the other hand, if the pyknons were not part of larger units, then the shifted strings would stride the original strings' "natural boundaries," and the number of their intergenic/intronic copies would change drastically. See Supporting Text for the results for pyknons in 3′ UTRs and separately for the intergenic and intronic regions; the curves for $d = 0$ correspond to the pyknons in $P_{3'\text{UTR}}$. Note that, even for a shift of $d = +2$, the derived new strings have strikingly fewer intergenic and intronic copies than the pyknons in $P_{3'\text{UTR}}$. We obtained similar results for negative values of d (data not shown).

The pyknons are possibly linked to PTGS. The most conspicuous feature of Fig. 2 is the preference for distances typically encountered in the context of PTGS. Recall that the 127,998 pyknons have one or more instances in the untranslated and coding regions of human genes: for each pyknon, we generated its reverse complement $\bar{\beta}$, identified all of $\bar{\beta}$'s intergenic and intronic instances, and, using the VIENNA package (40), predicted the RNA structure and folding energy of the immediately surrounding neighborhoods. We discarded structures that were predicted to self-hybridize locally or whose predicted folding energies were > -30 kcal/mol (1 kcal = 4.18 kJ). We also discarded structures that contained either a single large bulge or many unmatched bases. Each of the surviving regions was predicted to fold into a hairpin-shaped RNA structure that had a straightforward arm–loop–arm architecture, contained very small bulges, if any, and was energetically very stable. The analysis identified 380,084 nonoverlapping regions predicted to form hairpin-shaped structures (298,197 in intergenic and 81,887 in intronic sequences). These 380,084 regions contained instances of the reverse complement of 37,421 pyknons (29.24% of total). In terms of length, the majority of these regions are between 60 and 80 nucleotides long. See *Supporting Text* for information on each chromosome about the density of the surviving regions per 10,000

Table 1. Partial list of biological processes whose corresponding genes show significant enrichment (green cells) or depletion (red cells) in pyknon instances in their 5′ UTR, CR, or 3′ UTR

GO Term	5′ UTR		Coding		3′ UTR	
	ENRICHMENT DEPLETION	|log(*P* value)|	ENRICHMENT DEPLETION	|log(*P* value)|	ENRICHMENT DEPLETION	|log(*P* value)|
DNA catabolism		19.88				23.05
muscle cell differentiation		27.11		28.08		
regulation of transcription, DNA-dependent		10.88		5.49		27.11
regulation of physiological process		8.63		5.14		22.11
nucleo-(base,side,tide)& nucleic acid metabolism		11.57		46.78		10.27
regulation of metabolism		3.58		8.95		27.33
DNA transposition				229.02		13.86
DNA metabolism		2.84		155.71		8.61
DNA replication				154.68		7.30
morphogenesis		3.27		4.43		9.93
organogenesis		6.58		2.34		9.26
		2.33		10.88		34.41
		5.15		13.97		15.72
				15.84		11.86
		3.1		34.01		23.25

All log (*P* values) have been Bonferroni-corrected. Terms that are consistently enriched (resp. depleted) in all three regions are colored green (resp. red). The full set of entries and information on the color convention of the cells with log (*P* values) is listed in Table 4 in *Supporting Text*.

nucleotides. Recall that the typical pyknon length is similar to that of a microRNA and that there is a straightforward sense–antisense relationship between segments of the 380,084 hairpins and the pyknons instances in human 5′ UTRs/CRs/3′ UTRs. We also note that the 81,887 hairpins that originate in introns account for 21,727 of the 37,421 hairpin-linked pyknons and will be part of transcribed regions. If pyknons are, indeed, connected to PTGS, then Fig. 2 suggests that (*i*) in addition to 3′ UTRs, PTGS is likely effected through the 5′ UTRs and amino acid coding regions, and (*ii*) RNAi products in animals likely fall into distinct categories with preferences for lengths of 18, 22, 24, 26, 29, 30, and 31 nucleotides.

The pyknons relate to known microRNAs. We formed the union of the RNA family database Rfam (34) and pyknon collections and clustered it with a BLASTN-based scheme, using a threshold of pair-wise remaining sequence similarity of 70% (equals up to six mismatches in 22 nucleotides). When comparing two sequences A and B, we examined for redundancy the pairs (A,B) and (reverse-complement-of-A,B). In total, 1,087 known microRNAs clustered with 689 pyknons across 279 of the 32,994 formed clusters. See also *Supporting Text*.

The pyknons relate to recently discovered 3′ UTR motifs. We compared the pyknons in $P_{3'\mathrm{UTR}}$ to the 72 8-mer motifs that were recently reported to be conserved in human, mouse, rat, and dog 3′ UTRs (32). We say that one of these 8-mer motifs coincides with a pyknon of length ℓ if one of the following conditions holds: the 8-mer motif agrees with letters $\ell-7$ through ℓ of a pyknon ("type 0" agreement); the 8-mer motif agrees with letters $\ell-8$ through $\ell-1$ ("type 1" agreement); or the 8-mer motif agrees with letters $\ell-9$ through $\ell-2$ ("type 2" agreement). Of the 72 reported conserved 8-mer motifs, 39 were in type 0 agreement, 10 in type 1 agreement, and 7 in type 2 agreement with one or more pyknons from $P_{3'\mathrm{UTR}}$. Six of the 8-mer motifs did not match at all any of the pyknons in $P_{3'\mathrm{UTR}}$. In summary, the pyknons that we have derived by intragenomic analysis overlap with 56 of the 72 motifs that were discovered through cross-species comparisons.

Human pyknons are also present in other genomes, where they associate with similar biological processes. Table 2 shows, for each of seven genomes in turn, how many positions in region X of the genome at hand are covered by the human pyknons contained in set P_X, X = {5′ UTR, CR, 3′ UTR}. We account for length differences by

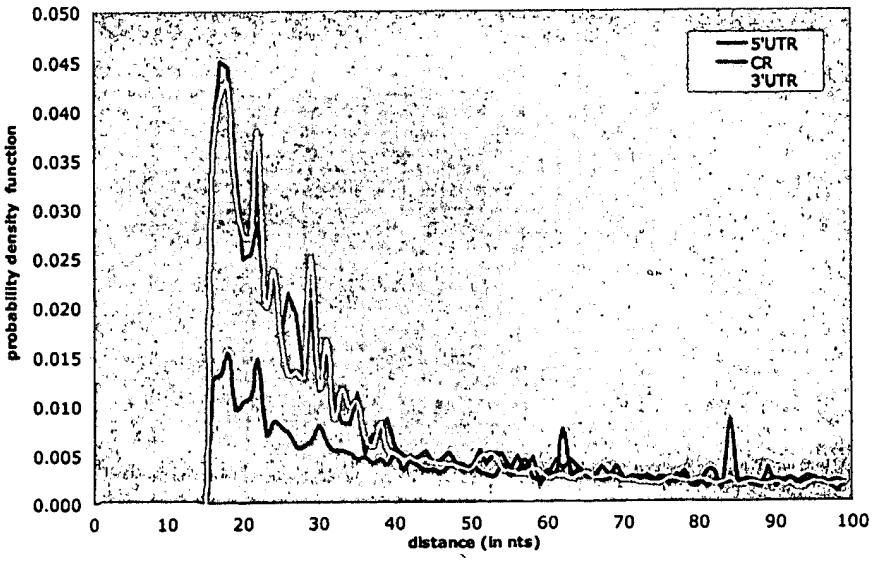


Fig. 2. Probability density functions for the distance between the starting points of consecutive instances of pyknons, shown separately for 5′ UTRs, CRs, and 3′ UTRs. The distributions have long tails, and only a portion is shown. Note the peaks at x = 18, 22, 24, 26, 29, 30, and 31.

Table 2. Number of positions per 10,000 nucleotides that are covered by instances of the human pyknons

Pattern set/ region	Positions covered in corresponding region of listed genome per 10,000 nucleotides						
	HSA	CFA	MMU	RNO	GGA	DME	CEL
$P_{5'UTR}$/5'UTR	382.2	43.0	20.7	14.2	9.7	22.4	5.8
P_{CR}/CR	304.1	88.3	57.4	61.1	28.4	25.2	14.7
$P_{3'UTR}$/3'UTR	733.3	152.4	82.6	64.9	42.1	65.9	57.1

For each of the three regions and each genome in turn we search region X of the genome with the patterns contained in the set P_X, where X = {5' UTR, CR, 3' UTR}. HSA, human; CFA, dog; MMU, mouse; RNO, rat; GGA, chicken; DME, fruit-fly; CEL, worm. See also text.

reporting the number of covered positions per 10,000 nucleotides. Table 3 shows how many of the human pyknons contained in set P_X are also present in the region X of the genome under consideration, X = {5'UTR,CR,3'UTR}. For each of the seven analyzed genomes, Table 3 also shows the number of intergenic and intronic positions covered by: (*i*) all human pyknons and (*ii*) those human pyknons that have instances in the corresponding genome's 5' UTRs/CRs/3' UTRs. Notably, >600 million nucleotides that are associated with nongenic copies of pyknons in the human genome are absent from the mouse and rat genomes. Interestingly, the human pyknons have many instances in the intergenic and intronic regions of the phylogenetically distant worm and fruit fly genomes, covering ≈1.6 million nucleotides in each.

A set of 6,160 human-genome-derived pyknons are simultaneously present in human and mouse 3' UTRs, whereas a second set of 388 pyknons are simultaneously present in human, mouse, and fruit fly 3' UTRs. Strikingly, we found these two sets of pyknons to be significantly overrepresented in the same biological processes in these other genomes (i.e., mouse and fruit fly) as in the human genome, even though the pyknons were initially discovered by processing the human genome in isolation (see Table 5, which is published as supporting information on the PNAS web site). The common processes include regulation of transcription, cell communication, signal transduction, etc. Finally, for each of the 388 pyknons in this second set, we manually analyzed ≈130 nucleotide-long neighborhoods centered on the instances of each pyknon across the human, mouse, and fruit fly 3' UTRs, for a total of >4,000 neighborhoods. Notably, we did not find any instance of syntenic conservation across the three genomes.

Discussion

We explored the existence of links between coding and noncoding sequences of the human genome and identified 127,998 pyknons with a combined 226,874 nonoverlapping instances in the 5' UTRs, CRs, or 3' UTRs of 30,675 human transcripts (20,059 genes). In transcripts that contained multiple pyknon instances, we were surprised to find the pyknons arranging themselves combinatorially, forming mosaics. Further analysis revealed that the UTRs and/or CRs of genes associated with specific biological processes are significantly enriched/depleted in pyknons.

We also found that the pyknon placement in 5' UTRs, CRs, and 3' UTRs is strongly biased: The starting positions of consecutive pyknons show a clear preference for distances between 18 and 31 nucleotides. Importantly, we found an apparent lack of correlation between consecutive pyknon instances in these regions. The observed bias in the relative placement of the pyknons is conspicuously reminiscent of lengths that are associated with small RNA molecules that induce PTGS, suggesting the hypothesis that the pyknons' instances in these regions correspond to binding sites for small RNAs. Analysis of the regions immediately surrounding the intergenic and intronic instances of the reverse complement of the 127,998 discovered pyknons revealed that 30.0% of the pyknons have instances within ≈400,000 distinct, nonoverlapping regions between 60 and 80 nucleotides in length that are predicted to fold into hairpin-shaped RNA secondary structures with folding energies ≤−30 kcal/mol. Many of these predicted hairpin-shaped structures are located inside known introns and, thus, will be part of transcribed regions. Our analysis also suggests that PTGS may be effected through the genes' 5' UTR and amino acid regions, in addition to their 3' UTRs. Another suggestion is that RNAi products in animals likely fall into distinct categories, with preferences for lengths of 18, 22, 24, 26, 29, 30, and 31 nucleotides. Notably, through sequence-based analysis, we showed that ≈40% of the known microRNAs are similar to 689 pyknons and that the pyknons subsume 56 of the 72 recently reported 3' UTR motifs, lending further support to the possibility of a connection between the pyknons and RNAi/PTGS.

The intergenic/intronic copies of the 127,998 pyknons constrain almost 900 million nucleotides of the human genome. Instances of human pyknons are also found in the nongenic and genic regions of the worm, fruit fly, chicken, mouse, rat, and dog genomes, and the numbers of found human pyknons decrease with phylogenetic distance. Strikingly, the human pyknons that we found inside the 3' UTRs of mouse and fruit fly were overrepresented in the same biological processes as in the human genome. We note that >600 million bases, which correspond to identically conserved intergenic/intronic copies of human pyknons, are not present in the mouse and rat genomes.

The fact that some of the intergenic/intronic copies of pyknons originate in repeat elements may lead one to assume that our

Table 3. Number of human pyknons that are conserved in the human genome and the corresponding region of the *j*th genome for seven genomes and for each of 5' UTR, CR, and 3' UTR

Genome	No. of human pyknons with instances in the corresponding region			Total size of interegenic/intronic region (both strands)	Intergenic/intronic positions (both strands) covered by	
	$P_{5'UTR}$	P_{CR}	$P_{3'UTR}$		all human pyknons	pyknons "in common"
HSA	12,267	54,396	67,544	6,093,304,675	692,393,548	692,393,548
MMU	400	8,767	6,160	5,216,777,897	89,568,584	45,996,326
RNO	170	3,424	1,644	5,409,179,291	82,635,080	25,134,158
CFA	234	6,170	1,351	4,826,002,769	87,572,989	7,912,193
GGA	51	1,786	718	1,855,717,211	9,262,198	577,232
DME	174	1,335	1,175	228,181,521	1,562,508	559,698
CEL	20	996	790	170,879,577	1,634,993	174,174

Shown is the number of intergenic/intronic positions in the *j*th genome that are covered by (*i*) all human pyknons, and (*ii*) only those human pyknons that are also present in the *j*th genome's 5' UTRs/CRs/3' UTRs. HSA, human; CFA, dog; MMU, mouse; RNO, rat; GGA, chicken; DME, fruit fly; CEL, worm. See also text.

analysis has merely "rediscovered" such elements. However, as mentioned above and in the *Supporting Text*, >50,000 of the pyknons have many of their instances in repeat-free regions. Moreover, the typical length of a pyknon is substantially smaller than, e.g., that of an Alu element. It was recently reported that genes can achieve evolutionary novelty through the "careful" incorporation of Alu elements in their coding regions (41, 42). Also, the "pack-mule" paradigm revealed that entire genes, large fragments from a single gene, or fragments from multiple genes can be "hijacked" by transposable elements (43). "Fortuitous coincidence" is generally considered the prevailing mechanism by which such potential is unleashed. In contrast to this view, the combinatorial arrangement of the pyknons within the untranslated and coding regions of genes, together with the large number of instances in these regions, their tight packing, and the association of pyknons with specific biological processes, suggests that their placement is not accidental and likely serves a specific purpose. Our findings do not rule out a link with transposable elements; instead, they seem to support a dynamic view of a genome (44) that has learned to respond, and likely continues to do so, to environmental challenges or "stress" in a controlled, organized manner.

The results of the analysis suggest the existence of an extensive link between the noncoding and gene-coding parts in animal genomes. It is conceivable that this link could be the result of integration into the genome of dsRNA-breakdown products. Because many genes are known to give rise to antisense transcripts, it is possible that these genes were, at some point, subjected to RNAi-mediated dsRNA breakdown, which, in turn, gave rise to products ≈20 nucleotides in length. The latter, through repeated integration, could have eventually given rise to the numerous intergenic and intronic copies of the pyknons that we have identified. However, this explanation would have to be reconciled with four of our findings. First, the pyknons have identically conserved copies in nongenic regions. Second, pyknons appear to favor a specific size and, in genic regions, a specific relative placement. Third, slight modification of the 3′ UTR instances of the pyknons, by either prepending or appending immediately neighboring positions, results in new strings whose intergenic and intronic copies are markedly decreased. And fourth, we can discover human pyknons in other organisms, such as the mouse and the fruit fly, where they exhibit a persistent enrichment within specific processes, yet are not the result of syntenic conservation. It may well be that we are seeing traces of an organized, coordinated activity that involves nearly all known genes. The existence of a pyknon-based regulatory layer that is massive in scope and extent, originates in the noncoding part of the genome, operates through the genes' UTRs and CRs, and is linked to PTGS is a tantalizing possibility. Moreover, the observed disparity in the number of intergenic/intronic positions covered by human pyknons in the human and the phylogenetically close mouse/rat genomes suggests that pyknons and, thus, the presumed regulatory layer, may be organism-specific to some degree ("pyknome"). Addressing such questions might eventually help explain the apparent lack of correlation between the number of amino acid coding genes in an organism and the organism's apparent complexity.

Methods

Under the assumption that all four nucleotides are independent and identically distributed, we estimate the probability p of a pattern of length l to be $P = 4^{-l}$. The probability Pr_k to observe k instances of a given pattern in a database of size D ($D \gg 1$) is then $Pr_k \approx (pD)^k e^{-pD}/k!$ (Poisson distribution). The least specific pattern that our method will discover is one that is the shortest possible (i.e., $l = L = 16$) and appears the fewest allowed number of times (i.e., $k = K = 40$). If $D = 6.0 \times 10^9$ bases (i.e., all chromosomes, both strands), then $Pr_k = 1.95 \times 10^{-43}$. In *Supporting Text*, we recalculate Pr_k using the nucleotides' natural probability of occurrence. Whether we assume equiprobable nucleotides or use their natural frequency of occurrence in our calculations, even the least specific pattern remains statistically significant. Alternatively, we can estimate the significance of our patterns using z scores: For the least specific patterns of length 16 with only 40 intergenic/intronic copies, we obtain the remarkably high value of $z = 32.66$; longer patterns and patterns with more copies have even higher z scores. These analyses separately confirm that every one of our discovered patterns is statistically significant and not the result of a random process. These conclusions hold true for the reverse complements of the discovered patterns and for the pyknons, the latter being a subset of the discovered patterns P_{init}.

We thank Annie Visviki, Laxmi Parida, Alan Grossfield, and the anonymous reviewers for comments and suggestions on the manuscript.

1. Mattick, J. S. (2004) *Nat. Rev. Genet.* **5**, 316–323.
2. Ruvkun, G. (2001) *Science* **294**, 797–799.
3. Ettwiller, L. M., Rung, J. & Birney, E. (2003) *Genome Res.* **13**, 883–895.
4. Brazma, A., Jonassen, I., Vilo, J. & Ukkonen, E. (1998) *Genome Res.* **8**, 1202–1215.
5. Lenhard, B., Sandelin, A., Mendoza, L., Engstrom, P., Jareborg, N. & Wasserman, W. W. (2003) *J. Biol.* **2**, 13.
6. Sinha, S. & Tompa, M. (2002) *Nucleic Acids Res.* **30**, 5549–5560.
7. Wasserman, W. W. & Sandelin, A. (2004) *Nat. Rev. Genet.* **5**, 276–287.
8. Hobert, O. (2004) *Trends Biochem. Sci.* **29**, 462–468.
9. Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S. & Haussler, D. (2004) *Science* **304**, 1321–1325.
10. Dubchak, I., Brudno, M., Loots, G. G., Pachter, L., Mayor, C., Rubin, E. M. & Frazer, K. A. (2000) *Genome Res.* **10**, 1304–1306.
11. Frazer, K. A., Sheehan, J. B., Stokowski, R. P., Chen, X., Hosseini, R., Cheng, J. F., Fodor, S. P., Cox, D. R. & Patil, N. (2001) *Genome Res.* **11**, 1651–1659.
12. Jareborg, N., Birney, E. & Durbin, R. (1999) *Genome Res.* **9**, 815–824.
13. Miziara, M. N., Riggs, P. K. & Amaral, M. E. (2004) *Genet. Mol. Res.* **3**, 465–473.
14. Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K. D., Ovcharenko, I., Pachter, L. & Rubin, E. M. (2003) *Science* **299**, 1391–1394.
15. Dermitzakis, E. T., Kirkness, E., Schwarz, S., Birney, E., Reymond, A. & Antonarakis, S. E. (2004) *Genome Res.* **14**, 852–859.
16. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. (2003) *Nature* **423**, 241–254.
17. Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W. & Lawrence, C. E. (2000) *Nat. Genet.* **26**, 225–228.
18. Brennecke, J., Hipfner, D. R., Stark, A., Russell, R. B. & Cohen, S. M. (2003) *Cell* **113**, 25–36.
19. Elbashir, S. M., Lendeckel, W. & Tuschl, T. (2001) *Genes Dev.* **15**, 188–200.
20. Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E. & Mello, C. C. (1998) *Nature* **391**, 806–811.
21. Johnston, R. J. & Hobert, O. (2003) *Nature* **426**, 845–849.
22. Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. (2001) *Science* **294**, 858–862.
23. Lee, R. C. & Ambros, V. (2001) *Science* **294**, 862–864.
24. Lim, L. P., Glasner, M. E., Yekta, S., Burge, C. B. & Bartel, D. P. (2003) *Science* **299**, 1540.
25. Moss, E. G., Lee, R. C. & Ambros, V. (1997) *Cell* **88**, 637–646.
26. Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelli, A. E., Bettinger, J. C., Rougvie, A. E., Horvitz, H. R. & Ruvkun, G. (2000) *Nature* **403**, 901–906.
27. Slack, F. J., Basson, M., Liu, Z., Ambros, V., Horvitz, H. R. & Ruvkun, G. (2000) *Mol. Cell* **5**, 659–669.
28. Ambros, V., Lee, R. C., Lavanway, A., Williams, P. T. & Jewell, D. (2003) *Curr. Biol.* **13**, 807–818.
29. Mattick, J. S. & Makunin, I. V. (2005) *Hum. Mol. Genet.* **14**, R121–R132.
30. Poy, M. N., Eliasson, L., Krutzfeldt, J., Kuwajima, S., Ma, X., Macdonald, P. E., Pfeffer, S., Tuschl, T., Rajewsky, N., Rorsman, P. & Stoffel, M. (2004) *Nature* **432**, 226–230.
31. Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H., Kelly, K., *et al.* (2005) *PLoS Biol.* **3**, e7.
32. Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S. & Kellis, M. (2005) *Nature* **434**, 338–345.
33. Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., *et al.* (2005) *Science* **308**, 1149–1154.
34. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S. R. (2003) *Nucleic Acids Res.* **31**, 439–441.
35. Rigoutsos, I. & Floratos, A. (1998) *Bioinformatics* **14**, 55–67.
36. Stabenau, A., McVicker, G., Melsopp, C., Proctor, G., Clamp, M. & Birney, E. (2004) *Genome Res.* **14**, 929–933.
37. Wootton, J. C. & Federhen, S. (1993) *Comput. Chem.* **17**, 149–163.
38. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215**, 403–410.
39. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., *et al.* (2000) *Nat. Genet.* **25**, 25–29.
40. Hofacker, I. L., Fontana, W., Stadler, P., Bonhoeffer, L. S., Tacker, M. & Schuster, P. (1994) *Monatsh. Chem.* **125**, 167–188.
41. Iwashita, S., Osada, N., Itoh, T., Sezaki, M., Oshima, K., Hashimoto, E., Kitagawa-Arita, Y., Takahashi, I., Masui, T., Hashimoto, K. & Makalowski, W. (2003) *Mol. Biol. Evol.* **20**, 1556–1563.
42. Lev-Maor, G., Sorek, R., Shomron, N. & Ast, G. (2003) *Science* **300**, 1288–1291.
43. Jiang, N., Bao, Z., Zhang, X., Eddy, S. R. & Wessler, S. R. (2004) *Nature* **431**, 569–573.
44. Jorgensen, R. A. (2004) *Cold Spring Harbor Symp. Quant. Biol.* **69**, 349–354.

ADDENDUM #7

Molecular Modelling Investigations of (i) Nucleic Acid Triplexes Comprising Nonisomorphous Base Triplets and (ii) PNA.DNA Duplexes Relevant to Antigene Strategy

THESIS

Submitted to the

UNIVERSITY OF MADRAS

in partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

by

R. THENMALARCHELVI

DEPARTMENT OF CRYSTALLOGRAPHY AND BIOPHYSICS

UNIVERSITY OF MADRAS

GUINDY CAMPUS

CHENNAI 600 025

INDIA

April 2004

CONTENTS

Part-I Structure and Dynamics of Nucleic Acid Triplexes

Chapter 1

Nucleic acid triplexes: An overview

Chapter 2

Methods

Chapter 3

Residual twist as a measure of base triplet nonisomorphism: Implication to sequence dependent nonuniform DNA triplex

Chapter 4

Influence of residual twist on the antiparallel DNA triplex comprising nonisomorphic G•GC and T•AT triplets: An alternating conformation for the RH strand

Chapter 5

Influence of residual twist on the antiparallel DNA triplex comprising nonisomorphic G*GC and A*AT triplet neighbourhood: A molecular dynamics investigation

Chapter 6

Influence of residual Hoogsteen twist on parallel DNA triplexes: MD simulation of a parallel DNA triplex comprising nonisomorphic G•GC and T•AT triplets

Chapter 7

Influence of a nonisomorphic base triplet interrupt in parallel DNA triple helices

Chapter 8

Modelling of antiparallel DNA triplexes formed at the base pair inversion site

Chapter 9

Potential of purine rich oligoribonucleotides to target DNA duplex for triplex formation: MD simulation of RNA•DNA.DNA (R•DD) hybrid triplexes

Part-II Molecular Dynamics Simulation of DNA.PNA and DNA Duplexes Formed at the Ki-*ras* Promoter

Chapter 10

Influence of base pair mismatch in DNA.PNA duplexes: Structure and dynamics of DNA.PNA duplexes formed at the Ki-ras promoter

Chapter 11

Effect of base pair mismatch in DNA duplexes formed at the Ki-ras promoter: Investigation by MD simulation

Appendix

A. Estimation of unperturbed end-to-end dimensions of 2',5' polynucleotide chain

Preface

The research work reported in the thesis has been carried out by the candidate in the Department of Crystallography and Biophysics, University of Madras, under the guidance of Dr. N. Yathindra, during the period 1998-2004.

The thesis consists of two parts. **Part-I** concerns with the elucidation of the influence of nonisomorphic base triplets on the fine structural aspects of nucleic acid triplexes that result from an interaction of single stranded sequence specific DNA and RNA oligonucleotides with a DNA duplex. It is proposed that a quantitative estimation of base triplet nonisomorphism may be made in terms of a pre-existing twist or a residual twist between the base triplets and, the influence of base triplet nonisomorphism on the triplex structure may be explained in terms of such pre-existing or residual twist. This has been demonstrated through *in silico* experiments by considering a number of DNA and RNA•DNA.DNA hybrid triplexes comprising nonisomorphic G•GC&T(U)•AT (parallel) and G•GC&A•AT (antiparallel) and G•GC&T(U)•AT (antiparallel) base triplets. One of the major outcomes of this study is that the residual twist may be responsible for sequence dependent nonuniform structural variations in DNA triplexes comprising nonisomorphic base triplets. **Part-II** examines the effect of A...C mismatch on the structure of DNA.PNA and DNA duplexes formed at the Ki-*ras* promoter. Molecular dynamics simulations carried out to investigate this have demonstrated that the presence of A...C mismatch has more destabilising effect in the DNA.PNA duplex than in the DNA duplex. It is argued that this perhaps be the reason for the experimentally observed less stable nature of DNA.PNA duplexes in the presence of a mismatch. The results obtained are expected to lead to a better understanding of the structure and dynamics of nucleic acid triplexes, DNA.PNA duplexes and, aid in the better design of antigene molecules for gene regulation.

Part-I

It is well known that nucleic acids assume a triple helical structure by accommodating a third oligonucleotide strand along the major groove of a Watson and Crick paired DNA duplex. Triplex Forming Oligonucleotides (TFOs) recognise the

purine rich strand of a DNA duplex by forming either a Hoogsteen or reverse Hoogsteen base pair. Thus, TFOs, interfere with RNA polymerase or transcription factors causing inhibition of gene expression. This is generally referred to as antigene strategy of gene regulation.

An overview of the structural aspects of nucleic acid triplexes forms the main content of **Chapter 1**. Triplexes may be formed by pyrimidine (C, T/U), purine (A, G) and purine rich (G, T/U) oligonucleotides. Pyrimidine TFOs interact with the purine strand of a DNA duplex in a parallel fashion, by forming isomorphic T•AT and C^+•GC triplets. Isomorphic or isosteric nature of these two triplets is expected to result in a "regular" or "uniform" structure for the triplex. This is, in a way, very similar to the Watson and Crick DNA duplex formed by isomorphic or isosteric A...T and G...C base pairs, especially, in the absence of contextual sequence effects. TFOs comprising G and T can interact with the purine strand of the DNA duplex, both in parallel and antiparallel orientation, by forming G•GC and T•AT triplets. However, a TFO comprising G and A is known to favour antiparallel orientation for its interaction with the purine strand of the target DNA duplex, by forming G•GC and A•AT triplets. A distinguishing feature in all of these triplexes is that the base triplets in them are nonisomorphic with one another, in sharp contrast to isomorphic T•AT and C^+•GC triplets. In view of this, a triplex structure formed by nonisomorphic base triplets is expected to be nonuniform, unlike the triplex formed by the isomorphic T•AT and C^+•GC triplets. This is to be anticipated even in the absence of possible stereoelectronic effects resulting from juxtaposition of base triplets. Nonisomorphic nature of base triplets although has been recognised in the literature, there has been no attempt either to characterise it or to define it. Nor there has been any attempt to carry out a systematic analysis to deduce their effect on the structure and conformation of DNA triplexes. The thesis mainly addresses these issues through *in silico* studies by considering a number of triplexes comprising a variety of nonisomorphic base triplets. The results obtained are expected to provide a comprehensive understanding of the influence of base triplet nonisomorphism on nucleic acid triplexes.

Molecular Mechanics (MM) and Molecular Dynamics (MD) simulations have been quite successful in elucidating the structure and dynamical aspects of

macromolecules. These methods are extensively employed in the investigation reported in the thesis. **Chapter 2** outlines the methods used in the *in silico* modelling investigations. Brief accounts of the force field and various optimisation techniques are also provided. Algorithms and various approximations used in the molecular dynamics (MD) simulations are indicated along with the protocols that are followed during the MD simulation and analysis of the trajectories.

Chapter 3 elucidates how nonisomorphism between a pair of base triplets becomes readily amenable for quantitative description. It is shown that, in general, a pair of nonisomorphic base triplets may be associated with a pre-existing twist between them. This pre-existing twist is referred to as the **intrinsic residual Hoogsteen twist (Δ°)**, in a parallel triplex where the Hoogsteen hydrogen bond scheme is used to form base triplets. It is referred to as the **intrinsic residual reverse Hoogsteen twist (Δ°)** in an antiparallel triplex where the reverse Hoogsteen hydrogen bond scheme is used to form base triplets. It is demonstrated that Δ can provide a convenient measure of base triplet nonisomorphism and, the degree or extent of nonisomorphism is relatable to the magnitude of the intrinsic residual twist Δ. It is also shown that the value of Δ is found to be 10.6° and 9.8° between the antiparallel G•GC and T•AT triplets and, G•GC and A•AT triplets respectively. It is found to be rather high (Δ=21.8°) between the parallel G•GC and T•AT triplets. This **Chapter** also discusses the effect of such residual twists in DNA triplexes consisting of alternating nonisomorphic (i) G•GC & T•AT antiparallel triplets (ii) G•GC & A•AT antiparallel triplets and (iii) G•GC & T•AT parallel triplets as revealed by molecular mechanics (MM) investigations. The results of such a study indicated that the intrinsic residual twist exerts a strong mechanistic influence leading to helical twist angle variations at the adjacent steps of the Hoogsteen and reverse Hoogsteen duplexes of the corresponding triplexes. The results thus provided a stereochemical basis for the sequence dependent DNA triplexes.

In order to obtain a greater insight into the effects of residual twist and base triplet nonisomorphism, MD simulation (4ns) has been carried out on just over one turn of (14mer) a 12-fold antiparallel DNA triplex comprising alternating G•GC and

T•AT nonisomorphic triplets. The results revealed a number of interesting and unexpected features and these are reported in **Chapter 4**. One of them relate to the observation of alternating high and low twist angles at the alternating GT and the TG steps respectively of the reverse Hoogsteen duplex. This is in line with the results of MM study (**Chapter 3**). Most interestingly, the bases in the third strand undergo significant changes to adopt an alternating *high anti* conformation for guanines and *anti* conformation for thymines. Another feature is the occurrence of concomitant alternating (g^-, g^-) and (t/g^+, g^-) phosphodiester conformations at the TG and GT steps respectively of the reverse Hoogsteen strand. Such alternating conformational features in the side chain and backbone result in a zigzag structure for the third strand. These could be directly attributed to the effects of Δ defining the base triplet nonisomorphism between G•GC and T•AT triplets. The results are compared with a lone NMR investigation on an intramolecular antiparallel DNA triplex comprising nonisomorphic G•GC and T•AT juxtaposition. Detailed account of structural variations along with the water and ion interactions with the triplex forms the content of **Chapter 4**.

Chapter 5 discusses the structural perturbation caused by the nonisomorphic G•GC and A•AT base triplets in a DNA triplex. The results of 4ns MD simulation show here also the presence of alternating high and low twists at the GA and AG steps respectively of the reverse Hoogsteen duplex. Further, as in the case of antiparallel DNA triplex formed with alternating G•GC and T•AT triplets, alternating backbone phosphodiester conformation for the third strand is observed. It is argued that the enhanced base stacking seen here might be responsible for the higher T_m found for these triplexes compared to those formed with G•GC and T•AT triplets. Overall, the results demonstrated the significant influence of Δ on the antiparallel DNA triplex.

Chapter 6 investigates the nature of influence of large value of intrinsic residual Hoogsteen twist (Δ=21°) that exists between the nonisomorphic G•GC and T•AT triplets in a parallel DNA triplex. Results of 4ns MD simulation indicate that large value of Δ tends to disrupt the canonical G...G Hoogsteen hydrogen bonds in G•GC triplets, necessitated considerable rearrangements in the triplex caused by high

value of Δ. Possibility of formation of energetically less favourable noncanonical G...G Hoogsteen hydrogen bonds, which results in lowering the value of Δ is also revealed. Less stable nature of canonical G...G Hoogsteen hydrogen bond seems to offer a stereochemical rationale for the unstable nature of parallel DNA triplex with nonisomorphic G•GC and T•AT triplets. Possible stabilisation of this triplex in the presence of counter ions is also discussed in the context of the strong coordination of metal ion suggested by the MD.

MD simulations (2ns) have also been carried out to assess the influence of a single G•GC and T•AT interrupts in a homopolymeric triplex. These further confirm the disruption of canonical G...G and T...A Hoogsteen pairs with the concomitant formation of noncanonical G...G and T...A Hoogsteen pairs. These indicate that the presence of even a single nonisomorphic base triplet causes destabilisation leading to local distortion in the DNA triplex. Possible stabilisation of noncanonical hydrogen bonds either through ion or water mediation also become evident under this situation. These results suggest that the interrupting nonisomorphic base triplets may be likened to a base triplet mismatch. Details of these results form the contents of **Chapter 7**.

Poly(purine).poly(pyrimidine) stretch in the genome sequences are often interrupted by one or more base pair inversions. When such inversions are centrally located, the poly(purine).poly(pyrimidine) sequences can be regarded as the sum of two abutting sites, each potentially capable of forming a triplex. In this connection, formation of triplexes at a critical 27bp poly(purine).poly(pyrimidine) sequence interrupted by two adjacent CG inversions located in *bcr* promoter has been examined to explore the extent of cooperativity at the triplex junction. Suitability of using two separate TFOs to target the two triplex forming sites at the promoter, instead of a single long TFO that spans the inversion site is examined. This has been carried out by constructing several triple helices using a number of 13mer and 14mer TFOs that will create a variety of base juxtapositions. Energy minimisation studies carried out for these DNA triplexes with various base juxtapositions at the triple helical junctions show lack of continuity of base stacking interactions at the base inversion sites. Further, the results seem to suggest that usage of two TFOs instead of one that spans

the base pair inversion sites may not significantly contribute towards stabilisation of the antiparallel DNA triplex. These results are compared with the experimental observations. Discussion pertaining to these forms the contents of **Chapter 8**.

Use of RNA strands as TFOs has an advantage in the sense that they can be endogenously generated. This helps in circumventing the problems regarding cellular delivery and endonuclease susceptibility. In this connection, many experimental studies have shown the ability of pyrimidine rich RNA TFOs to form a triplex with the DNA duplex. However, experimental studies on the formation of triplex using purine rich rTFOs are scanty. Among them, some support the formation of the triplex, while the others do not. It is in this context, MD simulations (5ns) have been carried out for antiparallel R•DD hybrid triplexes formed by the interaction of $r(AG)_7$ and $r(UG)_7$ TFOs with a DNA duplex to throw light on the stereochemical possibility of forming such hybrid triplexes. Results reveal large deformation in the reverse Hoogsteen hydrogen bonds in these triplexes, especially, at the termini. Although, the results are not as conclusive, they are clearly indicative of the less stable nature of R•DD hybrid triplexes formed by nonisomorphic base triplets. Likewise, MD (1ns) simulation on parallel R•DD triplexes formed using $r(UG)_7$ TFO shows that both G...G and U...A Hoogsteen hydrogen bonds get disengaged. This indicates that the large value of residual twist (Δ=21.8°) between the G•GC and U•AT triplets seems to have a pronounced destabilising effect in R•DD hybrid triplexes compared to its all DNA counterpart. These results are discussed in **Chapter 9**.

Part -II

In addition to conventional triplex-mediated transcription inhibition, PNAs are also known to downregulate gene expression by forming DNA.PNA hybrid duplex. An important advantage of this is that, it is not critical to have a purine rich stretch in a DNA duplex and, any sequence of the DNA can be a target. Also, PNA lacks formal charge on them and, hence it forms a more stable duplex than the corresponding DNA duplex. Surprisingly, certain mismatches drastically reduce the stability of DNA.PNA duplexes when compared to an all DNA duplex. In order to provide a stereochemical basis for these experimental observations, MD simulations (2.5ns) have been carried

out for a DNA.PNA duplex with and without a mismatch. For this purpose, a sequence corresponding to the Ki-*ras* promoter present in the pancreas cell, wherein one of the alleles is point mutated, is chosen. Interaction of designed PNA with the Ki-*ras* promoter leads to an A...C mismatch with the wild type allele and, a perfect Watson & Crick A...T base pair with the mutated allele. Results of these studies indicate that stacking loss is considerable between adjacent pyrimidines in the mismatched situation. Further, the fluctuating nature of A...C mismatch hydrogen bond is also evident from this investigation. These are suggested to be responsible for the lowering of T_m of DNA.PNA duplex, in the presence of mismatch. Details of the influence of mismatch on the structural property of DNA.PNA duplex come under **Chapter 10**.

In order to compare the effect of A...C mismatch in a DNA.PNA duplex and in a DNA duplex, MD simulations (2ns) have been carried out for the isosequential DNA duplexes. Unlike in a DNA.PNA duplex, stacking in a DNA duplex is retained, suggesting that the presence of A...C mismatch does not significantly reduce the stability of the DNA duplex. Different possible hydrogen bonding scheme for the A...C mismatch is also revealed here compared to the corresponding DNA.PNA duplex. The differences observed with respect to stacking and mismatch A...C hydrogen bond, are attributed to the differences in the topology of DNA.PNA and DNA duplexes. **Chapter 11** discusses the results of these investigations.

Appendix

Nature has used 3',5' linkages instead of 2',5' linkages to encode genetic information. Nonetheless, 2',5' linkages are sparingly used by nature in biological process. In order to provide an answer to this fundamental evolutionary question from a stereochemical perspective, results from this laboratory have shown that nucleic acids, even with 2',5' links, can indeed form duplexes with restricted flexibility for helical polymorphism. In this context, an inverse relationship with regard to the shape and dimension of repeating nucleotides and the type of linkage (2',5' vis-à-vis 3,5') is recognised. According to this, a preferred nucleotide repeat with C2'*endo* sugar pucker assumes a compact form (P...P = 5.9 Å) and, a preferred nucleotide repeat

with C3'*endo* sugar pucker assumes an extended form (P...P =7.5 Å) form in 2',5' nucleic acids. These are in sharp contrast to 3',5' nucleic acids where the reverse prevails. Statistical mechanical calculations have demonstrated that 3',5' polynucleotide chains with extended C2'*endo* nucleotide repeats lead to a higher unperturbed end-to-end dimension compared to the 3',5' polynucleotide chains with a compact C3'*endo* nucleotide repeat. Hence, an opposite trend may be expected in 2',5' polynucleotide chains in view of the above. In order to examine this, necessary mathematical formalism have been developed by invoking a three virtual bond scheme to account for the major conformational flexibility in a 2',5' linked polynucleotide chain. Results show that the extended C3'*endo* repeating units lead to higher end-to-end dimension than the compact C2'*endo* nucleotide repeats in 2',5' linked polynucleotide chain. This trend is opposite to that seen for 3',5' polynucleotide chains. These results are described in **Appendix**.

Publications arising out of this research work:

1) Xodo, L.E, **Rathinavelan, T.**, Quadrifoglio, F., Manzini, G., and Yathindra, N. (2001), Targeting neighbouring poly(purine.pyrimidine) sequences located in the human *bcr* promoter by triplex-forming oligonucleotides, *Eur. J. Biochem.*, **268**, 656-664.

2) **Thenmalarchelvi, R.** and Yathindra, N. (2003), "Residual" Hoogsteen twist as a measure of non-isomorphism in DNA base triplets: implications to sequence dependent DNA triplexes, *Recent Trends in Biophysical Research, A volume in honour of B.D. Nag Chaudhuri & S.N. Chatterjee, Eds. Maiti, M., Suresh Kumar, G. and Das, S.*, 17-24.

Manuscripts under preparation:

1) **Thenmalarchelvi, R.** and Yathindra, N., Influence of residual reverse Hoogsteen twist on the structure of antiparallel DNA triplexes: A MD investigation.

2) **Thenmalarchelvi, R.** and Yathindra, N., Is the large value of residual Hoogsteen twist responsible for the less stable nature of parallel DNA triplex comprising G•GC and T•AT triplets?

3) **Thenmalarchelvi, R.** and Yathindra, N., Molecular dynamics investigation on R•DD hybrid triplexes formed by purine rich rTFOs.

4) **Thenmalarchelvi, R.** and Yathindra, N., Molecular dynamics simulation to investigate the effect of A...C mismatch on the structural property of DNA.PNA and DNA duplexes formed at the Ki-*ras* promoter.